Summer 2007

Teachers' Education and Training Factors and Their Influence on Formative Assessment Processes

William Clark Reed
Old Dominion University

TEACHERS' EDUCATION AND TRAINING FACTORS

AND THEIR INFLUENCE ON FORMATIVE ASSESSMENT PROCESSES

by

William Clark Reed
B.S., June 1969, United States Naval Academy
M.S. Ed., August 1996, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

URBAN SERVICES

OLD DOMINION UNIVERSITY

July, 2007

Approved by:

John M. Ritz (Director)

William H. Graves, III (Member)

Steven Myran (Member)

Richard Strauss (Member)

ABSTRACT

TEACHERS' EDUCATION AND TRAINING FACTORS
AND THEIR INFLUENCE ON FORMATIVE ASSESSMENT PROCESSES

William C. Reed
Old Dominion University, 2007
Dissertation Chair: Dr. John M. Ritz

This study sought to identify factors in teachers' education and training that may be associated with their capacity to use formative testing to inform instruction and, ultimately, improve their students' achievement. This research involved the identification of teachers' educational and training variables that might influence their abilities to analyze formative testing results, interpret the analyses, and modify instruction so as to improve students' achievement in third grade mathematics. The goal was to identify those factors in (1) teachers' educational histories and (2) teachers' professional development and training histories that contribute to their capacity to use formative testing results to inform instruction. Data were collected from 46 teachers by interview and survey, existing records, and the expert opinions of school district coordinators. Collected data were subjected to principal component analysis (factor analysis) revealing three components as professional training, program design and analysis, and instructional planning. The professional training component represented participation in seven training topics including assessment, testing, evaluation, test results analyses, instructional planning and revision, and data-driven and differentiated instruction. This component appeared to address all aspects of using formative testing results to inform instruction. The design and analysis component included graduate level courses in research design, assessment, testing, and test analysis. The elements of this component appeared to offer the participant a framework and a detailed appreciation of why formative testing can

improve achievement. The final component, instructional planning, represented graduate level education in standards, advanced curriculum design, differentiated instruction, and evaluation. The content of these courses appeared to provide insight into the conversion of testing results into meaningful instruction based on those results. These components may provide insight into the topics in teacher graduate education and professional development that contribute to a teacher's capacity to successfully use formative testing results to inform instruction as realized by improved student achievement in mathematics.

This dissertation is dedicated to teachers who strive to open our minds and empower us to become lifelong learners and seekers, to become everything that we have the potential to become. Without their often selfless efforts, many students would not progress much beyond the circumstances to which they were born.

When we are born, our spirits slow to the tempo of earthly form. We inherit parents and are situated in families, communities, states, provinces, countries, regions, continents, cultures, and religions over which we exercised no preference. As we grow, we acquire the beliefs, values, moralities, ethics, knowledge, ways of thinking, and economic means associated with our families, not necessarily our potential. Our opportunities are normally commensurate with the prosperity, health, economics, focus, compassion, politics, station, and circumstances into which we were born. We are defined and entitled or restricted accordingly. The collective result of these and many other factors define our respective realities. For many, reality is taken for granted, an inheritance. All that is or is not afforded challenges some, but limits many. Often, there is no way for the individual to know otherwise, yet we are judged accordingly.

If we are blessed, however, we are given enlightened parents or the gift of inspired teachers who understand these realities. They are often able to breach our respective realities and allow us to realize our respective potentials. The inklings of other possibilities that they sow may, someday, allow us to unravel and understand the conditions of our realities and endow us with the desire to seek and understand the realities of others. Our parents, teachers, and mentors cannot do the thinking for us but rather they can imbue us with the skills, curiosity, and courage needed for our journeys.

In doing so, the teachers that so empower, endow, and imbue us are often imperiled by the collective realities that they would breach.

It is my hope that this study contributes to the ever changing thinking and administrative requirements of education that demand so much of a teacher's abilities, time, dedication, and inspiration, that those processes be understood, and, if possible, rendered into mechanisms that authentically assist and support the dedicated educator in order to allow them sufficient time to continually inspire their students.

William Clark Reed

ACKNOWLEDGEMENTS

I would like to acknowledge those individuals without whose assistance and support this project might not have been completed. Their encouragement and support was an essential component of my completing the task.

To my Maker, who gave me the ability to constantly grow, I give thanks. Through the many diverse opportunities provided, I have been taught a willingness to listen, a desire to change, and given the courage to endure.

To my committee, Dr. John Ritz, Dr. William Graves, Dr. Steve Myran, and Dr. Richard Strauss, I express my gratitude for your professionalism, patience, understanding, and general courtesy in guiding me to the completion of this project. May the tools that you have empowered me to use make you proud in the ensuing years.

To Dr. John Ritz, I thank you for staying vital to the process despite the departure and demise of committee members and a change of topic. Your guidance, contributions, and blacksmith-like fortitude have forged a strong melt and allowed this day to happen.

To Dr. William Graves, I thank you for your kindly encouragement and the opportunities that you so freely provided beyond all normal bounds. Your guidance has been thoughtful and mentoring.

To Dr. Steve Myran, I thank you for your collegial guidance, insights into the importance of the topic and its achievement. The moments of insightful camaraderie also put things in perspective.

To Dr. Richard Strauss, my assessment and local testing mentor and frequent intermediary and champion within the school district, my special thanks. Your ability to

define process and open doors is astounding. Without your patient assistance, this page might never have been written.

To Dr. Jack Robinson, your constant presence, guidance, and advice were critical in getting back on track after the loss of my first topic. Your sage counsel, advice, support, and belief in a student were elemental at a critical stage in this project.

To Dr. Sidney Vaughn, You have earned my appreciation for the opportunities to discuss issues of research, process, and analyses. Your assistance was always timely and illuminating.

To the memory of Dr. Wolfgang Pindar, a mentor lost along the way. His guidance, support, directness, and honesty always led me to new understanding and appreciation for reality.

And last, but by no means least, to my wife, Karen, whose constant belief in me was fundamental to my continuing efforts, even when the finances were bad, the hours long, and the mood dark. You have my eternal thanks, devotion, and availability for honey-do lists for all time.

William Clark Reed

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

An effective user of formative testing must integrate many skills. These include appreciation of standards-based curricula and instruction, understanding assessment, and, in particular, formative testing processes in support of the curricula, and the capacities to analyze and interpret test results (Johnston & Lawrence, 2004; McMillan, 2003; Thorndike, 1997), especially if ensuing instruction is to be differentiated. These elements are the heart of the formative testing process. Teachers' beliefs in assessment and testing processes are also possible considerations in the success of formative testing (Delandshere & Jones, 1999; Hughes, 1999; Black & Wiliam, 1998a, b; McMillan, 2001, 2003; McMillan, Myran, & Workman, 2002).

The skills necessary to effectively use formative testing are acquired from several sources. Contributors include the content and experiences of preparation programs, graduate education, classroom experience, professional development, and association with more knowledgeable and/or experienced colleagues. Continued development of teachers' capacities to accomplish formative testing tasks is also influenced by their school district's policies as well as their schools' environments and procedures. The goal of this research was to examine variables derived from education and training histories and identify factors potentially contributing to teachers' effective use of formative testing to inform instruction in third grade mathematics.

Statement of the Problem

The problem of this study was to examine the possible sources contributing to a teacher's capacity to use formative testing results to inform instruction. The capability of

interest is an amalgam of elements potentially influenced by a teacher's beliefs about the formative testing process. The elements of interest include the content and focus of teacher preparation programs, experience in the classroom, the focus and content of graduate programs, and selected professional development and training. These factors may lie beneath the effective use of formative testing to inform instruction in order to increase learner achievement in third grade mathematics.

Research Questions

As teachers develop the capacity to use formative testing to inform instruction, several sources are potentially reflected in the resulting skills. The individual teacher's capacity to use the formative testing process is a combination of ability; personal beliefs about learning, assessment, and testing; undergraduate, pre-service, or teacher preparation program content and focus; experience with assessment, curricula, and instruction in the classroom; post-graduate or continuing education content and focus; and professional development received from district, school, and peers (Johnston & Lawrence, 2004; Darling-Hammond, 1994a; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Assessment Reform Group, 2002; Ash & Levitt, 2003; Athanases & Achinstein, 2003; Popham, 1998; Delandshere, 1996; McMillan, 2003). Generally, these elements and the learning of teachers constitute variables that may be contributors to underlying constructs representing the capacity to effectively use formative testing to inform instruction, and, in this case, might be reflected in their learners' achievement in third grade mathematics.

Thus, the research questions (RQ) to be answered include:

RQ1: Are the focus and selected content of teacher preparation programs and graduate

education variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as directed and measured by quarterly assessment results?

RQ2: Are selected contents of professional development or in-service training programs variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as both directed and measured by quarterly assessment results?

Both research questions refer to selected contents as variables. In each case, the participation in and completion of specific coursework, courses, or topics constitute the majority of variables associated with these research questions. The following are elaborations of the specific content addressed by each research question.

A teacher's education is potentially comprised of two levels of exposure. The first is undergraduate or preparatory programs leading to initial licensure and constitutes the basis of entry level skills. The second is graduate or continuing education programs typically taken as part of an advanced degree program or for professional enrichment and/or advancement. Beliefs may represent the stimulus behind selected education. Undergraduate education variables of interest are comprised of the focus of a teacher's undergraduate or preparatory program and coursework that included assessment, testing, test item writing, data-driven instruction, and courses in mathematics taken in addition to program requirements.

Selected content variables associated with graduate studies include the focus of a teacher's graduate or on-going education program and, in particular, courses with content in evaluation, statistics, testing, test item construction, differentiated instruction, research

design, assessment, standards, analysis of results, advanced curriculum design, and mathematics taken in addition to program requirements. Graduate and on-going education and more experience are likely to produce greater foundational knowledge about assessment and testing and the part they play in instruction (Mislevy, 1996, as cited by Haladyna, 2004). Graduate education beyond a master's degree is also a possible contributing variable.

Professional development and in-service programs are typically specific training topics focused on achieving desired or prescribed goals. Selected content or topics include mathematics methods, assessment, testing, test results analysis, data-driven decision-making, action research, evaluation, differentiated instruction, instructional strategies, student activities and exercises, and instructional planning and revision. The number of years that a teacher has taught is a consideration in training. More teaching experience connotes more teacher-training in topics of interest.

Background and Significance

Public education in the United States appears to be in the age of standards and assessments (Bedwell, 2004). Standards are an essential part of describing desired goals while offering the detail by which their attainment might be measured. Assessment or testing is the means by which achievement of the desired standards is documented. The concepts of standards and assessments are nearly timeless (Glaser & Silver, 1994) though their combination to improve public education is of relatively recent origin (Taylor, 1994; Koretz & Barron, 1998). State-level standards have only proliferated during the past decade and now exist in every state (Rigney & Martineau, 2005). With the adoption of standards, summative standards-based testing inevitably became the backbone of

accountability in public education (Linn, 2000).

Darling-Hammond (1994a) asserts that the way in which assessment is to be used in the standards and assessments environment is critical to the movement's success. She contends that assessment must become part of both the teaching and learning processes. Baker (1994) suggests that the standards and assessments movement provides opportunities for new approaches to assessment. While summative assessment of learning has been the basis of grades, placement, selection, or accountability during much of recorded history, it is Black and Wiliam (1998a) who offer that formative assessment or assessment for learning is the concept that warrants exploration and expansion.

Given that formative assessment is one of the new approaches Baker (1994) foresaw, Darling-Hammond's 1994 assertions also offered that the quality and fairness of assessment, the appreciation of assessment's limitations, the equitable application of assessment in educational processes, and the education and training that teachers need to effectively use assessment to inform skilled and adaptive teaching are all critical factors (1994a). The latter observation, cast in the terms of formative assessment, suggests that it is the teacher's capacity to use formative assessment to inform instruction and to direct subsequent educational processes that are critical. Despite the seeming imperative, preparatory program courses addressing assessment and the basis for data-driven decision-making, as anecdotally suggested by Johnston and Lawrence (2005), have affected 10 percent or less of teachers. To that end, teacher preparation programs, post-graduate education, and available professional development within their districts and schools are critical to teachers' effective use of formative assessment.

There is a certain irony associated with formative assessment. It is not a new idea.

The concept of formative assessment has been the cornerstone of informed instruction for as long as teachers have interacted with students. Socrates, 2400 years ago, was renowned for his ability to lead learning through questioning, listening to his students' responses, effectively evaluating those responses, and thereby formulating the next instructive cycle through a new question (MacDonald-Ross, 1993). Most modern teachers endeavor to interact with students in much the same way as learner responses to stimuli generally guide the next instructional step (Stiggins, 2002). Used in the context of the standards and assessments movement, however, formative assessment often implies an examination based on standards, resulting in tangible evidence of learner achievement in the form of fixed or constructed responses, and, when properly analyzed, offers a source of information on which subsequent instruction might be based. McMillan (2003) suggests that it is a reflective decision-making process that considers available evidence that is essential. Formative testing is the aspect of assessment that is of interest in this study.

If standards and assessments are principles used to initiate improved achievement, as pointed out by Darling-Hammond (1994b) and others (O'Neil, 1993; Resnick, Nolan, & Resnick, 1995; Ravitch, 1996; Smith, et al., 1997a; Baker & Linn, 1998), formative testing to inform instruction may be a primary process. There are drawbacks, however. Formative testing to guide instruction constitutes a paradigm shift that requires new perspectives, beliefs, and skills. In such cases, there is a minimum requirement to establish a culture conducive to these new perspectives. Development of ensuing capacity takes time to create. The creation of a supportive infrastructure, identification or creation of appropriate testing materials, specification of how the process will function,

acquisition of sufficient technology to process the responses generated, and collective knowledge needed to make all aspects function in a formative rather than the more common summative way are minimum requirements. Most of these requirements then depend on each teacher's ability to convert data and reflect on its meaning in order to appropriately inform instruction. The teacher's capacity is the critical element given that adequate technological capacity exists to process high volumes of test responses.

Skills critical to the formative testing process include the capacity to appropriately analyze results of properly prepared tests, draw meaningful insights from those results, and revise planned instruction accordingly. These processes are most effective when individual content constructs and individual students constitute the levels of analyses but starting treatment with a given construct at the classroom level is reasonable. The next round of formative testing then measures progress while simultaneously being used to initiate modifications to the next phase of instruction. The exact processes that a teacher with access to detailed testing data would use for formative purposes are seldom discussed in literature. This is not surprising as literature suggests that these topics are very complex and are strongly linked to teachers' beliefs about assessment (McMillan, 2003). Success is thus clearly dependent on a teacher's understanding of and preparation to accomplish such tasks (Black, 1998; Darling-Hammond, 1994b; Linn, 1994b; Lissitz, 2005; McMillan, 2003; Popham, 1998; Smith & Yen, 2005; Stiggins, 2001b, 2002). As more school districts make greater use of local testing data, hopefully interest in formative testing and contributors to teachers' skills to accomplish such tasks should increase.

As the basis for examination, the sum of teachers' undergraduate and/or

preparation programs and their foci, teaching experience, graduate and/or on-going education and their foci, and professional development defined by district, school, and team training are key. Not to discount personal experiences, within these sources lie the sum total of preparation to accomplish the tasks associated with formative testing to inform instruction. Educational contributions may include courses or topics within courses discussing standards, assessment, testing, item construction, design, research concepts, evaluation, statistics, analyses, decision-making, instruction, curriculum development, differentiation, and the revision or modification of curriculum and instruction. In educational courses or topics, understanding and the 'why' of the content are often the goals. Such content provides a foundation on which experience and additional education and training may build greater understanding and proficiency. Professional development or training seeks to achieve a specific stated goal, is often more focused, and stresses the 'how' of accomplishing a task. Together, these two contributors to professional capability should produce the skills and understanding desired. Moreover, if a specific content area is involved, such as mathematics, then additional preparation or training in that content area beyond basic requirements should improve a teacher's capacity to aide learner achievement. For the purpose of visualizing these contributory sources, a graphic model of contributors is depicted in Figure 1. It suggests a relationship among the variables examined in this study.

The current importance of formative testing to inform instruction, which includes the capacity to use data to make instructional decisions for classes, groups, or individual students, is critical to the standards and assessments movement. In that regard, the significance of this study potentially contributes to the research base of standards and

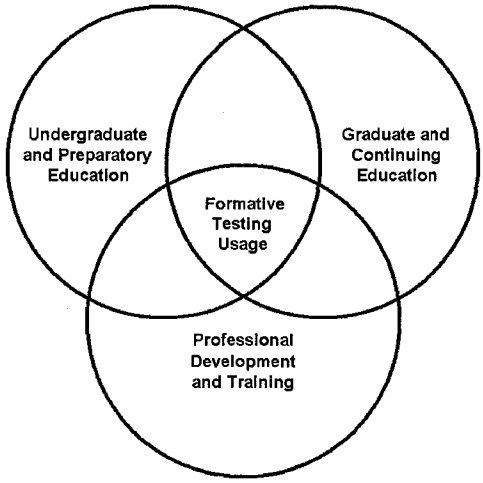

Figure 1. Tentative model of teachers' education and training influences on formative testing to inform instruction process

assessments in three ways. Regarding the nature of formative assessment, the study responds to Black and Wiliam's exhortation "to focus on the inside of the 'black box' and so to explore the potential of assessment..." (1998b, p. 8). This study examines one of the critical elements of the formative testing process black box: the contributors to a teacher's capacity to use testing results to guide instruction. The examination of influences that may help a teacher to achieve success using the formative concept is potentially contributory to better understanding formative assessment. The second potential contribution is the description of potential relationships between contributors to teachers' capacities made by undergraduate and graduate education and professional development and training. The third contribution potentially serves as evidence for examining the content of preparation, continuing education, and\or professional development of teachers with regard to formative testing. Given that standards and assessments remain a viable educational reform movement, a teacher's capacity to use formative testing to the learner's advantage will remain an ongoing area of developmental interest.

There is a technological significance to the study as well. Provided that the factors revealed enable teachers to effectively use formative testing, without the data required to fuel the process, the factors revealed are likely moot.

Setting

The school district in which this research was conducted is situated in the second poorest urban area in the Commonwealth of Virginia. Based on 2003-2004 school district performance reports, 2004 Commonwealth Report on Poverty, and 2000 Census Bureau data, the district is an intermediate sized city with 32.7%, or about one-third, of its elementary students eligible for Title I services (2003-2004 District Performance Reports). Title I percentages in individual schools range from 23% to 100% (2003-2004 School Performance Reports). Considering families with children under 17 years of age, 27.7% are at or below poverty levels. This percentage, however, represents the greatest total number of students (10,910) in any city or county within the Commonwealth of Virginia (U. S. Census Bureau, 2002). Median family income in the district is less than 74% of the national average and less than 68% of the Virginia average. The socioeconomic base of the community, as defined by the Free and Reduced Lunch (FRL) Program used in other studies (Wolf, Borko, McIver, & Elliott, 1999; Briars & Resnick, 2000; Kelley, Heneman, & Milanowski, 2000; Spillane, 2000; Betebenner, 2004) is substantial with 60.9% (Commonwealth averages 33.4%) of all students qualifying and represents the Commonwealth's greatest number at 22,400 students (VDOE, 2004). Between 2002 and 2004, this number rose five percent (VDOE, 2002; VDOE, 2004).

Faced with substantial achievement shortfalls – 47.1% passing the 1998 third grade Standards of Learning (SOL) Assessment in mathematics (2003-04 Division

Performance Report) – the formative testing program was established in an effort to improve achievement on SOL Assessments given in five content areas in grades three, five, and eight (VDOE, 1996). The adoption of the formative testing process as a means to affect achievement, especially with the gaps noted between ethnic groups, was seen as a diagnostic tool to guide instruction for otherwise disadvantaged children. While the local testing program is not the only strategy in force, formative testing is the only district effort that addresses academic processes in an objective, evidentiary manner. The stated goal of the program is to provide formative information to teachers (R. Strauss, personal communication, July 2003). The quarterly assessment program, accordingly, provides five content area tests every nine weeks, for grades three through eight, each based on the Virginia Standards of Learning as mapped-out in and paced by district curriculum and planning guides (R. Strauss, personal communication, July, 2003). Each of the district's thirty-five elementary and nine middle schools are required to participate. As of the 2003-2004 District Performance report, SOL achievement in third grade mathematics has risen to 81.4%, a gain of 34.3 percentage points.

For the local assessment program to be successful, teachers must be able to use the testing data generated to drive subsequent instruction and learning. If the factors that have contributed to this success can be verified, then there is real and general value in understanding each component. This is especially true if formative assessment, i.e., data-driven decisions to inform instruction, can be shown to affect critical sub-groupings such as Title I, low socioeconomic, African-American, and specific "at-risk" or special circumstance populations as these groupings constitute the most difficult Adequate Yearly Progress (AYP) hurdles in achieving mandated No Child Left Behind (NCLB)

criteria (Herman & Winters, 1994; Eisner, 2000; Washington Kids Count, 2001).

Limitations

The limitations of this study are as follows:

The aspects of testing program implementation are both complex and numerous and potentially confound the identification of factors that permit a teacher's conversion of testing results into informed instruction. Elements such as procedures for handling test results, analysis teams, or instructional planning teams could easily confound outcomes. Many of these aspects are products of a school's organization, support infrastructure, and personalities other than teachers.

This research seeks to quantify aspects of teacher's preparation and training that contribute to the capacity to use formative testing. Measurement of these variables are limited by the depth and accuracy of data received from teachers, coordinators, program managers, and records. As pointed out by many researchers (Spector, 1987; Bagozzi & Yi, 1990; Stecher, Barron, Kaganoff, & Goodwin, 1998; Stecher, Barron, Chun & Ross, 2000; Koretz & Berends, 2001), caution is appropriate when using self-reported data. While every effort was made to accurately and precisely collect data, details concerning courses taken, content experienced, and topics explored were subject to the quality of teacher recollections and the detail existent in records examined.

The actual content and focus of individual courses of interest in educational venues and the content and focus of specific training sessions are subject to variations related to a variety of institutions and instructors. These variations, without detailed specification, may invalidate the data collected and used in this research.

Participant teachers were selected from schools using a classic elementary school

teaching model, i.e., all core subjects taught by one teacher. This approach ignored potential contributions made to teachers' capacities by those who knowingly specialized in the teaching of mathematics and other specific content areas but less than all elementary core subjects.

Participant teachers were selected from schools using only district tests for school-wide testing. School-wide testing at a frequency greater than mandated by the district suggests a degree of organization, infrastructure, and processing that may, more likely, represent the school and are beyond the capacity of an individual teacher.

The selection of participant schools was based on the analysis of a single content area within an elementary school setting. The form of instruction (classic, all subjects versus departmentalized) and periodicity and scope of testing are of concern as findings may not be generalizable to other types of schools. While the basis of this study was mathematics, probably the most researched and best defined content area, the environment, conditions, and processes fostered by instruction in other content areas may be confounding.

Assumptions

Assumptions considered in this study are of concern as areas of interest adjacent to the focus of the study, such as instructional effectiveness, quality of district quarterly assessments, school organization, socioeconomics, etc., may confound results. The following assumptions are considered important:

Departmentalized instruction may motivate teachers to prepare more thoroughly to teach the subjects actually taught. This is opposed to the more balanced, generalist approach typical of more classic elementary education.

School-wide testing more frequent than prescribed by the district is likely to produce results affected by a higher degree of organization within the school and effects caused by more practice with the process of informing instruction not controlled for by the parameters of this study. Greater testing frequency also suggests that the processes might become more integrated into teaching practices and become less source specific.

The merits or flaws of the district's quarterly assessments are monitored by the program's coordinator and are considered fair and equivalent across the district. There should be little more than random effects on the consistency of outcomes.

District Quarterly Content Assessments are not high-stakes assessments but are intended to support learning and instructional improvement. The assessments are administered under generally standard conditions on prescribed dates and with integrity in that teachers monitor for cheating, prompting, or cuing of students. It is further assumed that teachers or schools have not engaged in any unethical practices which improperly prepare their students for an assessment. These perspectives are borne out in literature for low stakes testing (Haladyna, Nolen, & Haas, 1991; Koretz, McCaffery, & Hamilton, 2001).

Access to quarterly assessments, their blueprints, and their results is granted. The detail provided is sufficient to permit linkage of assessments by sub-standards on the provision that student, teacher, and school identity were protected.

Definition of Terms

For the purposes of this study, the following definitions of terms apply:

Assessment – in an educational setting, all activities used by teachers and students to measure learning achievement and to diagnose learning shortfalls (Black &

Wiliam, 1998b).

Assessment analysis – the process of reviewing testing results including item performance and associated statistical measures such as *p*-values; discrimination indices; reliability of items; a review for bias, skewing, or any effect that appears to bias or give unfair advantage to any member of the population subjected to the assessment; and an association of individual, group, or class performance to one or more constructs usually identified as a category, standard, or substandard.

Formative process – a cycle of instruction and assessment by which a teacher uses test results, analyzes those results for patterns and association with students and standards, and prescriptively modifies instruction in order to boost student achievement in those noted areas of weakness.

Formative testing – assessment, rendered as a test, providing evidence actually used to modify learning activities to meet student need (Black & Wiliam, 1998b).

Growth – long-term, within cohort, multiple measures, increase in achievement.

Informed instruction – instruction in which assessment is embedded thereby producing a "more skillful and adaptive teaching that enables more successful learning for all students" (Darling-Hammond, 1994a, p. 9).

Implementation plan – that combination of written or verbal policies and directives that: define objectives, functional elements, allocated resources, assignment of responsibilities, and conferral of authority; specify elements of accountability; establish procedures; set milestones; provide supportive training; and/or generally govern the accomplishment of the desired goal.

Improvement – the difference between any two status scores within a year and cohort.

In-service teacher – "refers to teachers who are currently teaching. Often used in the context of professional development: in-service teacher training" (Center for Research on Education, Diversity, and Excellence, 2002, p. 2).

Instructional amendment – the revision or change of planned instruction based on the diagnoses of learning achievement as suggested by assessment results analysis.

Professional development – "includes activities that- (i) improve and increase teachers' knowledge of the academic subjects the teachers teach, and enable teachers to become highly qualified; (ii) are an integral part of broad school-wide and district-wide educational improvement plans; (iii) give teachers, principals, and administrators the knowledge and skills to provide students with the opportunity to meet challenging state academic content standards and student academic achievement standards;". . . "(xiv) include instruction in the use of data and assessments to inform and instruct classroom practice" (Title IX, 20 U.S.C. 7801, section 9101 (34)(A), 2004).

Quarterly Content Assessment – assessments produced by the school district that comply with school district curriculum and pacing guides; guidance and input of the instructional department's subject matter experts; and the testing division's production, review, monitoring, and analysis efforts.

Standards of Learning – the Commonwealth of Virginia's published content and grade specific standards for achievement.

Status scores – the results of any given assessment or test.

Underlying mechanisms – unseen contents of a "black box" responsible for the effective function of the device. In this study, the *black box* of interest is the teacher and the

underlying mechanism is that mixture of a teacher's abilities, beliefs, education, training, and practices pertaining to formative testing used to inform instruction (Black & Wiliam, 1998b).

Summary

In this chapter, the concept of formative testing was introduced suggesting that its use required an integration of many educational concepts to be effective. It was further suggested that teachers acquired these elements from a variety of sources. The problem of interest was stated as an examination of the possible sources contributing to a teacher's capacity to use formative testing results to inform instruction. Two research questions were presented, one examining the formative testing process from the contributions made by undergraduate and graduate education and the second dealing with contributions made by topics presented through professional development and training. The contents and topics of interest were specified.

The background of formative testing was linked to the consensus that there was a standards and assessments-based reform movement underway in education. A brief discussion of the form that such a concept might take was presented suggesting that formative assessment was the concept that warranted exploration. It was also suggested by several contributors that it was a teacher's capacity to use formative assessment that was critical. Thus, preparation for the included tasks was critical to the movement's success. The fact that formative assessment was nothing new was suggested as was the need for a paradigm shift in educational thinking. The focal issue of the formative process was identified as the capacity to analyze and convert data into appropriate instruction. Educational and training contributors to the process were once again

elaborated and a model for their potential relationships presented. Appropriate technology, it was suggested, is critical to the support of the use of formative testing data. The setting for the study was specified as were the motivations and the stated goal of the subject school district's formative testing program. Practical limitations and assumptions about the study were specified and key concepts were defined.

Chapter II begins with a discussion of standards and assessments including the derivation of the formative testing for learning concept. A foundation is provided for topics that teachers should know in their efforts to use formative testing to inform instruction. The discussion begins with beliefs followed by rudimentary knowledge concerning standards, pacing, instruction and instructional planning, and assessments and formative testing. In the examination of the latter, understanding of items, test structures, the impact of administration and processing, and analysis are discussed leading to informing instruction. The chapter then briefly examines undergraduate, pre-service, graduate, and continuing education as well as professional development as contributors to the knowledge-base teachers should have. The chapter is summarized.

Chapter III presents the methodology and procedures to be used beginning with a discussion of the type of research conducted. The population of teachers from which the sample for this study was drawn is identified. Research variables are specified with a brief discussion of the types of data prescribed. Instrument design and derivation are revealed. The methods of data collection are identified as well as the field procedures in force. The statistical analysis to be used is identified as principal component analysis, a form of factor analysis, intended to reduce the total number of variables to a few relevant components and may lead to the identification of constructs underlying the ability to

transform formative testing results into informed instruction. The criteria for each aspect of the analysis are specified and cited. The chapter is summarized.

Chapter IV discusses the successes of the sampling strategy and data collection processes. The data analyses employed are discussed including potential limitations due to sample size. Preliminary analyses regarding each research area are presented including all significant findings. The results of factor analyses (principal component analyses) are presented for each research question. The results of the analyses are summarized.

Chapter V begins with a summary of findings and their possible meaning. Conclusions are presented. Based on the conclusion presented, recommendations regarding policy, practice, and research in the area of teacher education and training and the formative testing process are presented. Recommendations for future research are presented including suggestions for the modification of the current study, an expansion of the sample size, and a broadening of scope.

Chapter II

REVIEW OF RELATED LITERATURE

History is replete with efforts to improve education. Were all things possible, every learner would have an educator dedicated to his or her learning goals, sensitive to and flexible in response to his or her needs, alert to the opportunities presented each new day, aware of the student's moment-to-moment progress, and mindful of the next learning experience that should or could occur. Dedicated mentors have been invaluable to scholars throughout history. Unfortunately, public education in the twenty-first century is constrained by the availability of such masterful teachers, the inherent expense that such an approach would entail, and the nagging question of whether the learner's education, despite extraordinary instruction, met legislated minimum standards. This is the challenge of modern public education. There are standards, the varied influences of the communities in which learning is to occur, limited resources, and wide variability in teacher expertise and student abilities. A starting point begins with the first-generation state-level standards that now exist in all states specifying the minimum of what all public school students should know within that state (Rigney & Martineau, 2005).

State-level standards have proliferated in the past ten years. In Virginia, the Standards of Learning (SOL) were implemented by the Virginia Board of Education in 1995 (White, Sturtevant, & Dunlap, 2003) based on legislation dating to 1981 (Kittock & Sargent, 1995). The past ten years have witnessed alignment of district curricula to the SOL, the establishment of SOL Assessments (VDOE, 1998), and national legislation (Public Law 107-110, 2002 – No Child Left Behind – NCLB) that establishes goals for all students making achievement of those goals an accountability issue for school districts

and states. None of these efforts are particularly sensitive to the differences and impediments that exist among regions, school districts, or communities in which they are applied. Failure to achieve connotes sanctions (Linn, 2000; Lewis, 2001; Linn, Baker, & Betebenner, 2002). For those communities with the means, achieving prescribed goals has been a task of reallocation and focus. For those without the means, gaps that were relatively well-documented before state standards and assessments legislation (Massell, 1998; Catwali, 2003) became more threatening as issues to be eliminated. For districts with gaps, extraordinary efforts have been necessary to *catch-up*.

This chapter will review the literature related to these issues with a focus on formative assessment and testing and, in particular, what may constitute a teacher's capacity to use formative testing to inform instruction thereby improving learner achievement. Formative assessment and testing are concepts that, with the assistance of properly prepared teachers, provide ways for school districts to *catch-up* (Black & Wiliam, 1998b; Stiggins, 2002) and are particularly effective in the case of low achievers (Black & Wiliam, 1998b). In such cases, Black and Wiliam report effect sizes of 0.4 to 0.7 in formative assessment experiments. The employment of periodic, district-wide, formative tests by faculty empowered to use their results to inform instruction is heavily dependent on the capacity of teachers to accomplish the included tasks. The formative testing process is complex, multi-faceted, and requires substantial skill and will to be properly employed.

The chapter begins with a brief overview of assessment and the complexity of formative testing in the era of standards and assessments. It leads into a brief examination of the capacity to inform instruction using formative testing. The process or *things* that

teachers should know to make appropriate use of formative testing to inform instruction will then be briefly discussed. The examination of issues will continue into an exploration of where required skills might be acquired. It is suggested that the development of needed skills has several contributors. Each contributor influences, molds, and/or develops an individual's personal abilities and beliefs ultimately resulting in a teacher attentive to the best practices and concepts of the day. Undergraduate education is normally the first contributor. Likewise, programs intended to prepare persons from other disciplines for their first incursions into the classroom are also important. As many teachers pursue graduate or continuing education, this constitutes the second group of contributors. Finally, participation in professional development within their school districts, training within their schools, and association with more learned associates are also contributory. The chapter will end with a summary of the capacity to productively use formative testing, analyze results, make data-driven decision, and differentiate the results for classes, groups, and individuals so as to inform instruction (Darling-Hammond, 1994a). It is from these accumulated skills and their contributors that factors impacting the use of formative assessment should emerge.

Standards and Assessments

Standards and assessments as a means to affect educational reform have been a topic of discussion for decades (Linn, 2000). It is ironic that recent interest in published standards seems to have begun with assessment. Citing Kilpatrick (1992), Glaser and Silver (1994) suggest that the proliferation of achievement testing began with secondary education as early as 1845. Citing Engelhart (1950), by 1900, some school districts required written examinations for promotion from grade to grade. Glaser and Silver also

suggest that Thorndike's treatise on educational measurement (1904) had a large impact on the refinement and rise of assessment. The outbreak of World War I found the United States' armed forces using intelligence tests for the selection and placement of recruits (Crocker & Algina, 1986; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001). The observation that what any two individuals knew, despite transcripts and diplomas, could be substantially different is significant. By World War II, the armed forces had developed the means to assess individuals for proper placement within their ranks. During the 1950s, Thorndike suggests that educational testing became big business and was often used for hiring and promotion decisions (1997). Thorndike identifies educational measurement and assessments being used to make decisions related to instruction, curriculum, selection, placement, and personal awareness. The use of achievement testing for placement and selection continues today. In recent years, however, the use of assessment for teaching and learning is on the rise (Delandshere & Petrosky, 1998).

In addition to the study of assessment for the measurement of achievement, teachers' knowledge of assessment practices (Taylor, 1994; Glaser & Silver, 1994; Black & Wiliam, 1998a, 1998b; Stiggins, 2001a, 2001b, 2002), and assessment effects on classroom practices (Stecher, Barron, Borko, & Wolf, 1997; Stecher, Barron, Kaganoff, & Goodwin, 1998; Koretz, McCaffery, & Hamilton, 2001), the debate of assessment broadened to other related educational issues. Other topics included the assessment of teachers (Delandshere, 1996; Moss, 1996), financial resources (Baker & Linn, 1997), and accountability (Linn, 1998) to name a few. The need for change in classroom assessment was evident. Glaser and Silver (1994), quoting a 1979 National Institute of Education conference report, clearly indicate that testing, as practiced, was extraneous to instruction

and not helpful to teachers.

Linn (2001) suggests that interest in assessment as a means to reform public education contributed to the development of standards on which those assessments might be based. The irony that assessment led to standards begins with evidence that the concept of standards has existed for at least 2000 years (Madaus & O'Dwyer, 1999). The idea of definitively measuring the achievement of a skill requires that the skill be clearly defined. Assessments without standards seem meaningless (Krathwohl, Bloom, & Maisa, 1964; Berk, 1980). In the late 20th century, the concept of standards, first in the form of minimum competencies (Cunningham, 1986), then implemented on broader scales at state and national levels have become compelling forces (Bedwell, 2004; Haladyna, 2004; Hamilton et al., 2003). With the 1989 publication of the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics*, the first national level content standards became available. States were also in the process of creating academic standards. In Virginia, the Standards of Learning (SOL) were adopted in 1996 and, in 1998, SOL Assessments commenced. Unfortunately, state level standards and assessments remain a summative process continuing to support the concerns stated by the National Institute of Education, in 1979, that such assessments are of little help to teachers (Glaser & Silver, 1994).

Given that linkage of standards and assessments is both natural and necessary, assessment seems to have taken two different directions. As the literature re-enforces, many notable and praiseworthy research projects, past and present, contribute to the understanding of assessment primarily in two ways. These may best be summarized by the phrases *assessment of learning* and *assessment for learning* (Assessment Reform

Group, 1999). The former is known as summative assessment. Such assessment has recently regained prominence as an accountability mechanism testing the adequacy of public education as emphasized in public law (Baker & Linn, 1997; Crocker, 2003). With the advent of the NCLB Act, interest in summative assessment in public education is clearly oriented toward accountability while continued eligibility for much needed public funding is contingent upon compliance and success. A result is interest in assessment of learning at every opportunity or the prediction of success on "high stakes" assessments as being possible through all testing. This supports the suggestion of Black and Wiliam (1998b) that there is tension between summative and formative processes. As a basis for guiding learning, however, the latter is gaining more prominence. Formative assessment has been a necessity since skills worthy of being accurately replicated were taught though it has not been as evident or well-developed as its summative kin. That assessment must be formative or informative in nature and prescriptive of the next learning or instructional step is gaining momentum (Bass & Glaser, 2004).

Assessment of Learning

Whether used for of-learning or for-learning purposes, achievement of standards cannot be properly determined without some form of measurement. Clear standards are then foundational to both types of assessments (Yoon & Resnick, 1998). In that light, Virginia's adoption of the Standards of Learning in 1996 was a necessary precursor to launching state-wide assessments in 1998 intended, primarily, for accountability purposes. SOL Assessments are intended to determine attainment of minimum competencies (DeMary, 2005a). In that regard, the proliferation of state standards and assessments, like the SOL and SOL Assessments, have become the darlings of state governments

(Linn, 2000, citing Madaus, 1985) as state assessment programs are relatively inexpensive compared to other forms of educational improvement, can be mandated from a state capital, can be rapidly implemented, and have visible results (Linn, 2000).

With the proliferation of state standards and assessment programs, many researchers were called upon to examine their impact. Koretz, Mitchell, Barron, and Keith's report on the Maryland School Performance Assessment Program (1996), for example, suggests that the success of state-wide assessment programs in achieving desired reforms was in no way guaranteed but was dependent on the behaviors of educators in the classroom. Change in those behaviors was required, would be difficult to achieve, and would require professional development in order to facilitate transition to the standards and assessments perspective. Koretz's team also reports that there are concerns about the level of detail included in standards and the curriculum on which assessments were based. They recommend more investigation. Their comments were similar to the concerns that led to the publishing of *Taxonomy of Educational Objectives* (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). Standards need sufficient detail in both content and cognitive tasking to unambiguously specify what is expected. The statement of detail, however, poses a problem for some. Detailed standards are often emphasized as discrete objectives while those not specified, however integral, are ignored. The bigger problem is that standards not assessed are often ignored in the classroom. This is the basis, it is believed (Madaus, 1988; Smith et al., 1997; Linn, 1998; Koretz & Barron, 1998; Schmoker & Marzano, 1999; McTighe & Thomas, 2003), for a narrowing of curricula.

Assessment for Learning

While the expansion of state level, "high stakes" assessment programs continued, some educators focused on assessment that provided alternatives. As Baker (1994) suggests, the standards and assessments movement provided opportunities for new approaches to assessment. Darling-Hammond's (1994b) point that the way in which assessments were to be used is critical to the argument. She contends that assessment has to be integrated with both teaching and learning processes which is to intimate that it was not common at the time of her writing. She further states that the success of standards and assessments-based frameworks require quality and fairness, the realization that assessment has limitations, that assessment must be embedded in such a way that provides equal resources and opportunities for all learners, and that teachers must be trained in the effective use of assessment to inform skilled and adaptive teaching. The latter point was one of the guiding considerations in this research. It, as later echoed by Koretz et al. (1996), constitutes a paradigm shift or a change in the culture of education from the summative use of assessment for grading, sorting, screening, selecting, and promoting learners (Shepard, 2000a) to the formative use for assisting students to learn.

Formative assessment is the means by which success might most effectively be achieved in the standards and assessments movement. Black and Wiliam's offering that formative assessment or assessment for learning is the concept that warrants exploration and expansion (1998a) was timely and often cited. Black and Wiliam (1998b) stated that formative assessment was the sum of all activities that provide information or feedback to modify the learning process. Their contention was that formative assessment is on-going and its strength lay in a systematic and never ceasing stream of information. Gronlund

(1998) suggests that the constant assessment of success or failure should result in adjusting instruction. Anderson and Krathwohl more succinctly characterize it as "in-flight instructional modifications" (2001, p. 102). Optimally employed, formative assessment is a combination of constructive feedback from all assessment or testing including those that are external, those conducted by a teacher, and those of learners empowered with the capacity to self-assess. In these, collectively, resides assessment for learning.

To be clear, Stiggins (2002) aptly states, "Assessment for learning is about far more than testing" (p. 761). Gronlund (1998) asserts that the purpose of assessment is measuring all aspects of a unit of instruction, then, using the findings to improve learning. It includes every exercise, question, exploration, or examination in which what is known is compared to what is required. Of interest to this study is that aspect of assessment for learning that uses standards-based instruction coordinated with standards-based testing for the generation of formative information that might be used to prescribe a way ahead. Testing, when well designed, creates tangible evidence that represents what learners know. However, given that every effort is expended to align testing to standards and instruction and cognitive tasking levels and numerous other variables that could confound outcomes are compensated for, testing-for-learning, can be achieved through most properly constructed testing vehicles. It is in the use made of the results that the intention of the assessment-for-learning process may reside (Darling-Hammond, 1994a).

If formative assessment and testing are mechanisms of a successful standards and assessments movement, then the feedback that they generate is exceedingly important. Feedback that suggests to teachers an appropriate course of instruction and that informs

students of their successes and failures and how to correct the latter are valued. The value of feedback in training is well established (Sleight, 1993). It is an essential element of workplace regimens. For example, trainers have long known that more immediate feedback is more effective. Application to the world of education seems to be impeded by the implication that the process is behaviorist suggesting that it does not require higher order thinking. Ebel and Frisbie (1986) suggest that formative feedback concerning learner performance, as in the workplace case, needs to be as immediate as practical and take the form of modified instruction. Their explanation of feedback is based in Glaser's Basic Teaching Model (BTM) (1962). The Basic Teaching Model, depicted in Figure 2, suggests the flow of instruction and the relationship of feedback to each component of the teaching process. Feedback is also associated with the learner (Crocker & Algina, 1986). Gronlund and Linn (1990) emphasize the importance of continuous feedback to students in order to reinforce successful learning. Such views sound much like B. F. Skinner's operant conditioning (Skinner, 1963) but Moss suggests that the "testee" should expect a "rehash" (1992, p. 246) of what was successful and what was not with suggestions of how to improve his/her less than successful efforts. The nature of the feedback should be non-judgmental suggests Costa (1993). Schafer (1993) attributes to Stiggins (1991) and

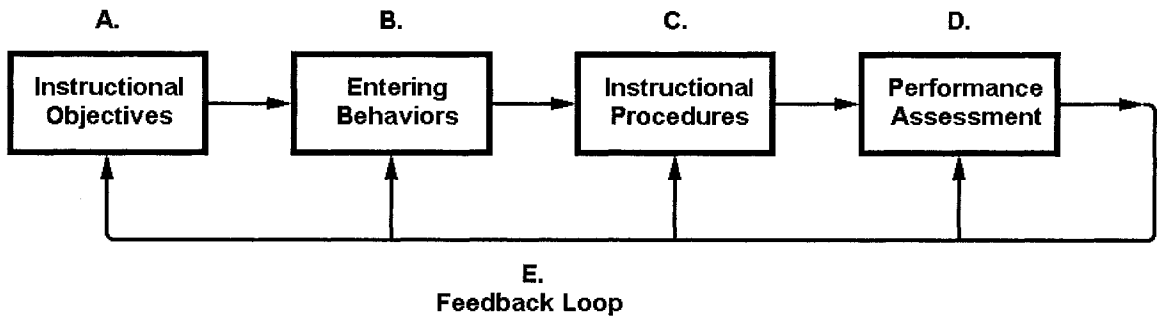

Figure 2. Basic Teaching Model (Glaser, 1962).

Schafer (1991) the idea that instruction concerning feedback from assessment is a critical part of teacher education.

The onset of the standards and assessments age recognizes that feedback is critical to both teacher and student. Taylor (1994) suggests that it is in a standards-based model that cycles of feedback and revision are essential. Wiggins (1994) maintains that the quality of feedback is only as good as the assessments generating it and that the quality and records of assessments and feedback must be maintained in a longitudinal fashion if they are to be of any value. Wiliam and Black (1996), citing Sadler's (1989) explanation of Ramaprasad's (1983) definition of feedback, reiterate that it only qualifies as feedback if it alters the gap between standard and status. Black and Wiliam (1998b) clearly stress the importance of feedback as opposed to grades reflecting on research in which the mere affixing of a grade to an assessment diminished its formative value. The description of formative feedback used by these and others is that it must be for an individual, constructively identify the learning goal, suggest the means to achieve it, and be immediate.

Formative assessment feedback must also, Black (1998) suggests, be constantly driven by the awareness of the differences between the learning goal and the learner's status. These "gaps" (Ramaprasad, 1983; Sadler, 1989) should be the targets of instruction, and further assessment and feedback. Black (2000) recommends Vygotsky's (1962) Zone of Proximal Development as a mechanism for managing the gaps. He contends that students who know what the gap is can readily manage its closure. The students who do not know what the gap is can be given a clear picture of what is expected and some ideas of how the differences might be overcome as they will eventually be

capable of closing their own gaps. Ravitz (2002) reemphasizes that the information about gaps or differences must be timely enough to help make a difference.

Teacher conducted formative assessment should be on-going to be effective. As cited earlier, Black and Wiliam (1998b) suggest that the effectiveness of formative feedback lies in systematic analysis that never ceases. Gronlund (1998) adds that the need to constantly assess learner successes and failures is the key to adjusting instruction. He further states that details as to who suffered which shortfalls are critical as re-teaching on a small group versus classroom basis or on an individual rather than a small group basis is fundamental to the formative process. The strategies that result should be articulated on an objective by objective basis. Gronlund elaborates suggesting that the most focused form of formative assessment is the diagnostic assessment which identifies common sources of error in a manner that defines corrective action. Gronlund's observations seem to support a relationship between formative feedback and the differentiation of instruction. Black and Wiliam (1998b) contend that allowing a student to analyze and work through their own weaknesses is considered critical. Confronted with tangible evidence of their difficulties constitutes the ultimate teachable moment, a moment in which the learner's own efforts are emphasized. Black and Wiliam's observations also seem to support differentiation.

Regarding the content of feedback, Stiggins (2002) argues that emphasis must be given in the form of descriptive rather than judgmental feedback. If an assessment is to be used formatively, without penalty, the process is likely to be trusted and more supportive of student learning. This argument supports Black and Wiliam's (1998b) suggestion that summative and formative assessments are at odds. Citing research in which the

substantive feedback was the same but adding a simple grade was perceived to be judgmental and compromised the formative value to the point of being worthless. Their contention that overemphasis of grades ameliorates the inclusion of meaningful feedback and that once existent, the included formative content is ignored. A grade, they suggest, is more likely to be the basis for comparison than for personal improvement.

Finally, the link of formative assessment to self-assessment is critical to Black and Wiliam (1998a). Their contention is that formative assessment establishes the pattern for self-assessment and learning from one's own mistakes and shortcomings. This skill is one that continues to be of use throughout life. To a student, self-assessment is possible if the goals are clearly articulated and the means to close the gaps are identified.

In spite of the merit attributed to formative assessment and the number of years in which formative assessment has been discussed, summative assessment appears to remain the favored method. In their 2003 report, McNair et al. suggest that despite efforts to be more formatively oriented, 76% of the third and fourth grade teachers encountered in their study continued to use their observations for summative rather than formative purposes. This is consistent with McMillan, Myran and Workman's (2002) findings for elementary school teachers and Black et al.'s (2004) report revisiting the Black Box. The latter authors stated that assessment measures in use still did not promote learning, still favored competition as opposed to improvement, and that less descriptive feedback was still having a negative impact especially with low achieving students. Johnston and Lawrence's (2005) anecdotal suggestion that ten percent or less of teacher preparatory programs study assessment and data-driven decision-making are likely responsible for the current weaknesses in assessment for learning practices.

The lack of progress with formative assessment, Black suggests, is attributed to the weak development of formative practices and external testing pressures (2000). He makes the case that summative practices continue to dominate the assessment landscape and continue to serve as poor examples of useful assessment. Moreover, summative assessments drive accountability. The reality is that most external assessments, such as SOL assessments, are limited in or devoid of sufficient detail to permit meaningful feedback. Reports of strand scores, or combinations of whole families of standards, are included with SOL Assessment reports, though linkage to specific items is not revealed. High-stakes tests are also infrequently administered. On the other hand, district-wide assessments on a quarterly basis with reporting to the sub-SOL and individual student levels may provide the detail required. This is a concept that seems consistent with both Marzano (2003) and Stiggins (2005).

In addressing these problems, it is worthy to note that most research of the late 1990s suggested that several things are necessary for assessment to be an effective agent of reform. Assessments have to be aligned to standards and curricula. Assessment results have to be used properly. Most importantly, the in-service training of educators must help them make the transition from a summative grades-oriented paradigm to a formative, for learning, paradigm. The differences in and uses of formative assessment remain a significant issue. As Shepard (2000) states, the shift requires a cultural change in the classroom. That professional development is a critical component of assessment, as stated by Ebel and Frisbie (1986), appears to continue. In the age of standards and assessments, teacher preparation and professional development have become even more important components in the success of reform efforts. Table 1 lists representative examples of

research conducted and reported in these areas of study. Specifics of these discussions suggest that standards are critical to making formative assessment goals achievable as is the need to shift from summative to formative perspectives.

Table 1
Representative Research Projects Conducted in Areas of Teacher Preparation and Professional Development

Teacher Preparation	Professional Development
Briars and Resnick, 2000	Darling-Hammond, 1994b, 2004
Crocker, 2003	Delandshere, 1996
Haertel, 1999	Koretz, McCaffery, Klein, Bell, and Stecher, 1993
Hamilton, McCaffery, Stecher, Klein, Robyn and Bugliari, 2003	Resnick and Harwell, 2000
Johnston & Lawrence, 2004	Simmons and Resnick, 1993
Koretz, Stecher, Klein, McCaffery, and Deibert, 1993	Stecher, Barron, Kaganoff, and Goodwin, 1998
McMillan, 2003	Yoon and Resnick, 1998
Popham, 1998	
Schafer, 1993	
Shepard, 2000a, b	
Stecher, Barron, Kaganoff, and Goodwin, 1998	
Stiggins, 2001b, 2002, 2004	

To achieve the desired goals, Marzano (2003) identifies five school level factors necessary for success. First, the curriculum must be the same no matter who is teaching and is taught at the time allotted. This factor suggests standardization in the form of and adherence to district curricula and pacing guides. Second, effective assessment (testing) should occur at least every nine weeks. With this approach, schools then have the data although they may lack the capacity to systematically monitor and guide student progress. This point supports the idea that periodic, district-wide, professionally developed tests have virtue if teachers adhere to published timelines. It is possible, however, that scheduled tests are potentially out of sync with individual classrooms. Third, parents and

the community are part of the effort to plan what the cycle of assessments entails. Fourth, an environment that is safe and orderly is essential. And, finally, teachers must be involved in the governance and be benefactors of a professional development program that addresses assessment and instructional strategies and take advantage of action research and assessment.

Thus, the use of the standards and assessments as a framework (Gronlund & Linn, 1990; Baker, O'Neil, & Linn, 1993; Baker, 1994; Darling-Hammond, 1994a, b; Linn, 2000; Marzano, 2003) for district level formative assessments administered multiple times a year could be of great value in realizing educational reform and closing the gaps in less affluent communities. The pivotal issue is whether teachers possess the required skills and capacities to use them. The district assessment process is especially valuable if it promotes the impartial diagnosis and identification of those constructs in which non-achievement is recorded and to which schools, classes, groups, or individuals such findings apply. Properly used, local, formative testing results should more closely reflect the state of actual achievement in schools than state level testing.

Information Teachers Should Know about Formative Testing

In this section, the literature regarding knowledge, concepts, and skills teachers should possess concerning assessment, and in particular, formative testing to inform instruction, will be examined. As an overarching perspective, Black and Wiliam (1998b) suggest, when considering teachers and assessment, that teachers simply do not have the time to adopt the "general principles" (p. 8) suggested by formative assessment nor the education and/or training to incorporate them. They explain that teachers' "classroom lives are too busy and too fragile for all but an outstanding few" (p. 8) to do the work

necessary to transition principles into practice. This observation suggests that skills to be acquired might have an appropriate conceptual basis but would be better presented in the form of tangible, how to knowledge. Risk is also involved. There is a risk associated with trying anything new. Failure may occur for no other reason than something is different. Unfortunately, failure deters many. Formative testing and improved measurement practices in the classroom may be perceived as risky endeavors. McMillan (2003) states that, despite the emphasis on classroom assessment and its link to instruction, teachers' daily practices remain inconsistent with best practices. Where fixed and selected response assessments are used, for example, simple item statistics, error measurements, and reliability checks, he suggests, are seldom if ever generated nor are sufficiently detailed test specifications or blueprints used in their creation. Instead, McMillan reports, there continues to be a high reliance on constructed response assessments with their inherently more subjective nature. In the era of standards and assessments, these are but two relevant aspects when considering what teachers need to know and do in order to effectively use formative testing. This brief examination of components of a formative testing to inform instruction cycle includes beliefs, standards, instruction, curriculum and pacing guides, test construction, test administration and processing, results analysis, and data use to inform instruction.

Beliefs

Beliefs underlie all human endeavors. One's beliefs can make the impossible possible or the possible impossible. Beliefs determine how one sees the world, whether those beliefs are valid or not. Succinctly put, "a belief is that which an individual holds to be true" (Smith & Shepard, 1988, p. 308). In *Taxonomy of Educational Objectives, The*

Classification of Educational Goals, Handbook II: Affective Domain, Krathwohl, Bloom, and Maisa (1964) consider beliefs a component of the affective domain of knowledge. Beliefs, the authors state, are built-up from experience. If new experiences or exposures are congruent with current beliefs, there is a tendency to accept them. This point is stated in another way by Smith and Shepard, i.e., "Beliefs are like emotional attitudes in that one can believe a proposition without realizing it" (1988, p. 308). To frame the importance of beliefs, McMillan (2003) associates many attributes of expert teachers to their personal beliefs and their appreciation of learning theories. If a teacher's role is finding effective ways of challenging a student's beliefs and getting them to consider expansions or alternatives, then teachers must also moderate their beliefs with theory, research, and continued learning.

As all teachers are first students, their experiences prior to entering a teacher preparation program are likely fundamental to their beliefs about teaching. Hollingsworth (1989) refers to these as "preprogram beliefs" (p. 161). Gerges (2001) states that pre-service teachers' beliefs about instructional practices, classroom management, and other aspects of education are formed by their own educational experiences. Pre-service coursework should then cause examination and adjustment of beliefs as necessary. Beliefs about what public education is supposed to accomplish and how effective it is (Cunningham, 1986), beliefs about the capacity of students to learn (Black & Wiliam, 1998a, 1998b; Hiebert, 1999; Jones & Vesilind, 1995; Stecher, Barron, Borko, & Wolf, 1997; Taylor, 1994), beliefs regarding the teacher or student-centeredness of learning (Prawat, 1992), or beliefs about assessment and testing are appropriate topics. Acceptance of new beliefs is possibly predictive of future success with the learning

processes. Ebel suggests that "It is good to hold beliefs that are reasonable. It is better to be able to cite empirical evidence in support of them" (1968, p. 321). To that end, pre-service teacher education is also about examining the "empirical evidence" in support of them. Ebel's goal is appropriate but Block and Burns (1976) contend "that it is beliefs and not data that run schools" (p. 41). Their observation, while dated, may remain unchanged. Swan (2006, citing Pajares, 1992), suggests beliefs must be a target for any institution attempting to develop teaching practices.

A principle area that requires work for most teachers is their fundamental beliefs about learning and their students' capacity to learn (Black & Wiliam, 1998a; Bliem & Davinroy, 1997; Delandshere & Jones, 1999; McMillan, 2003; Richardson, 2003; Shepard, 2000a). Briefly, if teacher beliefs are, as summarized from Black and Wiliam (1998a), that students have untapped potential as compared to a limited capacity to learn, he/she then assumes a perspective that biases his/her approach to teaching and appreciation of student learning. The teacher's assumptions then guide subsequent actions as evidenced in his/her assessment of and feedback to students. Alternatively, the pivotal belief may be that of teacher-centered (transmission) or learner-centered (discourse) educational model. A teacher's beliefs determine her/his instructional and assessment practices that lead to deep understanding, self-worth, and appreciation of the power of quality feedback versus the promotion of superficial learning where grades are emphasized over learning. Delandshere and Jones (1999) suggest that if learning is perceived as the accumulation of facts and skills, then assessment is seen through more summative eyes. Black (1998) suggests that teachers who emphasize the quality of work rather than the quantity often elicit higher achievement.

Beliefs about testing and the use of testing results to influence instruction are of interest here and have a substantial history. Haertel and Herman (2005) suggest that the concept of using testing results to drive instruction existed prior to the 1950s beginning with Tyler (1949). Tyler promoted a framework of instruction in which both curriculum and assessment were based on objectives. The authors attribute to Tyler the basis for today's formative assessment efforts, although there were a number of interim variations. Criterion-referenced testing, the authors go on to say, was a variation that was to be used in qualitatively different ways, examining student performance against criterion (standards) as opposed to normalized processes used for relative ranking or selection. They offer that these models may have evolved into the use of testing to drive individualized instruction.

The concepts of using testing to drive instruction have been the topic of much review and discussion. Shepard's (1991) contention that the use of testing aligned to curricula goals, then teaching the content to be tested or "teaching to the test" (p. 2), was a widely held psychometrician's belief and an inappropriate basis for learning and assessment. The ensuing dialogue supported by prior and subsequent works (Airasian, 1988; Bracey, 1987; Cizek, 1993; Frederiksen, 1994; Herman, 1997; Madaus, 1988; Noble & Smith, 1994; Popham et al., 1985; Popham, 1987) examined the concepts involved from many perspectives. The contention that using testing results to guide instruction was behavioristic in nature and was inconsistent with the rising constructivist view of learning seems to have been the death knell for, or at least a substantial barrier to, the concept. As there were substantial exchanges between Shepard and others based on their perspectives, it is safe to say that each argument and the research used to support

respective positions proceeded from each researcher's beliefs. As Barlow and Reddish (2006) citing Cooney and Wiegel (2003) suggest, beliefs act as influences or filters on what is understood and achievable. Because of beliefs, it may actually have been difficult for those espousing their perspectives to reconcile their positions into a unified concept that examined or made use of the best of all.

It is safe to say that beliefs are a potent force in determining what is achievable. What pre-service teachers learn, suggests Darling-Hammond and Ball (1998), is affected by their beliefs. Delandshere and Jones (1999) suggest that teachers' beliefs about assessment are then modified by the influences of external assessments, perceptions of curricula and subject matter, and understanding of learners and the learning process. What teachers practice is also affected by their beliefs. Black and Wiliam (1998a) suggest that a tension is created at the juncture beliefs and that which is imposed on teachers by external forces. McMillan suggests that it is this tension that explains why teachers use assessment as they do. His observation that beliefs, hence practices, "were not directly concerned with measurement principles" (2003, p. 36) suggests beliefs are critical to testing and its effectiveness even when they are not tied to the science of measurement. Altogether, the control of beliefs may be the secret to the success of formative testing to inform instruction but they are only the first of many necessary elements.

Standards, Pacing, Instruction, and Assessment and Formative Testing

The alignment of objectives, curriculum, instruction, learning, and assessment is a frequently discussed relationship in the literature (Cizek, 1993; Smith et al., 1997; Stecher et al., 1997; Neill, 1997; Pelligrino, Baxter, & Glaser, 1999; Herman, Brown, & Baker, 2000; Shepard, 2000a,b; Anderson & Krathwohl, 2001; Buckendahl, Impara, &

Plake, 2002; Haladyna, 2004). In a post-standards adoption era, curricula and objectives are usually founded on standards. If curricula, objectives, and instruction are aligned to standards, and assessment and testing are aligned to standards, then proper timing of assessment and instruction is a necessity in closing the alignment loop.

Standards

Fundamental to any valid educational or training program is the standards on which it is based. Standards, Reigeluth (1997) offers, serve two primary purposes. First, standardization specifies what all high school graduates should know. The second pertains to the level of attentiveness to a student's needs. A level of care seems to be connoted. In this study, the first definition is of interest. In the case of the Commonwealth of Virginia, standards were first called for in early 1980s legislation (Kittock & Sargent, 1995). The effort to employ standards was realized in 1996 as the Virginia Standards of Learning (SOL) were finalized and became operational. They soon became the bases of what was to be learned and assessed by SOL assessments in 1998. To achieve SOL assessment readiness, the formative testing to inform instruction process requires the acceptance of a system of standards that links instruction and assessment by providing enough specification to know exactly what is being taught and subsequently tested. It is with this thought in mind that an understanding of standards and, hopefully, standardization that follows are required teacher knowledge. The knowledge of standards seems to be assumed. As it is the basis for the whole process, assumption is not a sound enough foundation, especially in light of ubiquitous anecdotal evidence. The importance of standards and standardization of content area curricula and the cognitive tasks required cannot be understated. They are a foundational component of formative testing.

For most educators, standards are, like the SOL mandate, things to be achieved. Since standardization of knowledge is a primary goal, any influence that affects the interpretation of a standard affects the process of achieving it whether by formative or summative means (Berk, 1980). Interpretation begins with the inference of expectation articulated by authors of standards. In Virginia, SOL committees represented diverse groups of interested and expert parties. Their understanding and intentions are the basis for the meaning of the standards articulated. In the process of publication and distribution, meaning and possible conflict with the interpretation of meaning were inevitable.

Berk (1980) suggests that ambiguity permits multiple interpretations. The implementers, i.e., curriculum specialists and teachers, are faced with some degree of interpretation based on their respective understanding of standards and content areas as well as filtering by their fundamental beliefs. As Schmoker and Marzano (1999) suggest, there should also be concern about standards not assessed. In many cases, standards are judiciously amended on district levels in order to provide more detail or clarity. These alternatives do not omit published standards but rearticulate or add to them providing the necessary substance to make them useful. This process is referred to by Reeves as "adding value to standards" (2002, p. 8).

Additional standards evoke another of Schmoker and Marzano's suggestions: standards should not be more numerous than can actually be taught. Marzano (2000) suggests that standards must be explicit. Explicit standards clearly define what is expected and allow little room for interpretation. Anderson and Krathwohl (2001) support this position stating that specificity is a cornerstone of standards. Standards must then be comprehensive or supported by companion interpretations, otherwise a teacher with

insufficient content knowledge or expertise may not appreciate the relevance of underlying, integral, or necessary though unspecified content. The unspecified content represents Schmoker and Marzano's (1999) standards not assessed. Unless the standards are comprehensive and specific enough, relevant, subsumed, or understood knowledge may not be evident to or equally addressed by all teachers. This is a problem as Marzano (2000), citing an American Federation of Teachers report (reference not specified), suggests that only 13 of 49 state standards examined were specific enough to actually be used as guidelines for classroom activities.

The problem of interpretation is a frailty that teachers need to understand. Virginia SOLs have been revised and improved several times since their first publication. In some districts, the SOLs are considered inadequate as additional objectives are added by local curricula coordinators or committees as suggested by Reeves (2002). Additions generally evolve from the belief that something is missing or improperly articulated. These actions would suggest that the published state standards are incomplete or insufficiently detailed. In the era of standards and assessments, when standards from several states are compared – as has been done by a number of commercial companies seeking to vend automated assessment services – the levels of standards' detail are different enough to require the deconstruction of each state's standards into more basic segments or building blocks (NCS Pearson Measurement, 2005). What one state considers a standard might well be defined and articulated differently by another. Until a standard is reduced to its most basic, definable elements, what the standard actually means is open to interpretations. Interpretation threatens the value of standards.

The issue of standards is yet incomplete. Cognitive tasking of objectives or standards is a long-term issue, one notably articulated in Bloom, Engelhart, Hill, Furst, and Krathwohl's *Taxonomy of Educational Objectives, the classification of educational goals, Handbook I: cognitive domain* (1956). Cognitive tasking levels represented by standards require specification or multiple specifications depending on expectations. That each standard or included construct should be expressed in a number of cognitive ways is often unappreciated and abets interpretation. Unspecified, the expression of content constructs of interest is often realized in the least demanding way. Standards or objectives usually address content considered important but, according to most research in the area of standards, levels of cognitive effort are seldom articulated. Table 2 lists specific terms used when discussing cognitive effort and the research reporting on those cognitive processes. Without taxonomic consideration such as those suggested by the systems and levels listed in Table 3, the content is addressed but the depth of the learner's intellectual effort or ability to apply, analyze, create, evaluate, or communicate is ambiguously defined. While most credible works on testing refer to a two-way table, that is constructs defining rows and cognitive tasks defining columns (Gronlund, 1998; Notar, 2004), few

Table 2
Terms Used to Describe Levels of Cognitive Effort

Term	Source
Cognitive domain	Bloom, Engelhart, Frust, Hill, and Krathwohl, 1956
Cognitive task	Shulman and Elstein, 1975; Shepard, 1980; Ebel, 1982; Nickerson, 1989; Mislevy, 1996; Anderson and Krathwohl, 2001; Haladyna, 2004
Cognitive demand	Darling-Hammond, 1994a; Resnick, 1994; Mislevy, 1996; Klein, O'Neil, and Baker, 1998; Linn, 2001; Anderson and Krathwohl, 2001; Delandshere, 2002
Cognitive challenge	Borko, Stecher, Alonzo, Moncure, and McClam, 2003; Spielmann and Radnofsky, 2001; Stiggins, 2001a, b; Sloane and Kelly, 2003
Cognitive burden	Cizek, Bunch, and Koons, 2004

Table 3
Taxonomic Systems for Identification of Cognitive Effort

Taxonomic System	Reference
Bloom's Taxonomy	Bloom, Engelhart, Frust, and Krathwohl, 1956
Dimensions of Learning	Marzano, 1992
New Taxonomy of Educational Objectives	Marzano, 2001
Revised Bloom's Taxonomy	Anderson and Krathwohl, 2001
Families of Cognitive Demand	Baker, 2003

standards reflect them on state or district levels. For teachers, the importance of being aware of the cognitive function required of their students for a given construct or sub-standard cannot be understated.

In summary, teacher familiarity with the concepts and purposes of standards as defining and driving forces of instruction, appropriate and meaningful interpretation of those standards, and awareness of differences in cognitive tasking levels should be fundamental knowledge. Time permitting, teachers might also be introduced to the confounding influences of text readability, cultural, regional or language influences, and biases of any kind. Given that standards are specific enough, suggest pertinent level(s) of cognitive tasking, and are understood by the teacher, teachers using formative testing will likely have a sufficient basis to use results to inform instruction.

Pacing

If instruction and assessments are aligned with standards, the missing connections may be the sequencing and pacing necessary to permit prior knowledge to be built upon in a coordinated fashion (Marzano, 2003). Pacing can be considered in many ways. On an individual basis, differentiated instruction requires that a pace that fits the individual's needs is appropriate (Barr & Dreeben, 1977). Wiliam and Black (1996) citing Dahll f (1971) suggest that when teaching to a whole class, some teachers use a "reference

group" (p. 538) to determine the progress of instruction. In many districts today, curriculum guides are annotated with sequence and pacing information in a timeline manner that encourages a consistent approach to instruction across the district. While such guides have probably existed for decades, in one form or another, the Council of Great City Schools now consistently recommends that pacing be contained in curriculum guides (Richmond, 2003; District of Columbia, 2004). Pacing intervals include years, quarters, or months though annual intervals may be too ambiguous to be effective. The reality is that individual schools or teachers often elect their level of compliance with district policies as evidenced by efforts to monitor progress (Cohen & Ball, 1999; Linn, 2000; Ravitch & Brooking, 1998; Resnick, Nolan, & Resnick, 1995; Shepard, 2000). The potential impact of this variable compliance should be understood.

When it comes to state, district, or school-wide assessment efforts, non-compliance by any given teacher constitutes a threat to the validity and value of each assessment taken by his/her students. A virtue of these external assessments, Shepard (2001) offers, is the encouragement of uniformity in the pacing of instruction. Without some adherence to district pacing guidance, the potential for variation among classrooms could be substantial thereby invalidating any testing effort generated outside of a given classroom. Moreover, in low socioeconomic school districts, intra-district transfers could find students substantially ahead or behind the schools from which they transferred adding an academic burden to the existent economic situation. In such cases, the challenge for analysts, teachers, schools, or districts is the differentiation of results ascribing potentially poor performance to learner or institution (i.e., teacher, school, or district policies and pacing). With standards, objectives, sequence, and pacing guidance

in hand, the next challenge is planning for and instructing learners.

Instruction

From the student's perspective, report Black and Wiliam (1998b), it is the teacher who sets the stage that leads a learner to failure or success. This primarily occurs during instruction. It is during instruction that the culture of learning is established. If a teacher focuses on comparison of students and their achievements, it will feed the fears of some students that may eventually lead to their failure, especially in low achievers. In the preferred culture of success, the authors continue, formative assessment could change the tide by allowing each student to concentrate on his/her own specific problems by having clear understandings of the goal he/she fell short of and given alternatives how he/she might close the gap. With guidance of this kind, understanding the process of formative assessment to inform instruction allows for the eventual transference of assessment from the teacher to the self-assessment of the learner. To McMillan (2003), involvement of the student in the assessment process is critical. The ability to transfer responsibility takes education and skill development on the part of any teacher especially in a formative environment.

As a prelude to discussing instruction, given that beliefs are attended to and that standards exist and are understood, formative testing to inform instruction is a process which suggests a constant, dynamic cycle of activity. Wiliam and Black (1996) simplified the assessment process suggesting it was about eliciting evidence followed by construct-referenced interpretation. But these, they contend, are not formative unless they affect a closing of a learning gap. A little more complex perhaps, Almond, Steinberg, and Mislevy (2003) offer the Four Principle Processes in the Assessment Cycle model.

During planning, the teacher considers standards when formulating instruction, drawing on existent resources, and selecting appropriate activities. The instruction is presented to learners in appropriate ways. The learner is assessed, his/her responses, whether observed, verbal, or written, are processed and used to recommend further instruction. At all times, the four processes are linked to the body of standards, objectives, methods, and resources necessary for producing evidence of student progress. Feedback, critical to all elements of the cycle, occurs as required or desired. Almond et al.'s model is presented in Figure 3. In the case of formative testing to inform instruction, the emphasis is on renewing instruction based on high quality testing at every opportunity. Testing is one of the assessment options that produces tangible evidence. This cyclic nature of formative process is held by many.

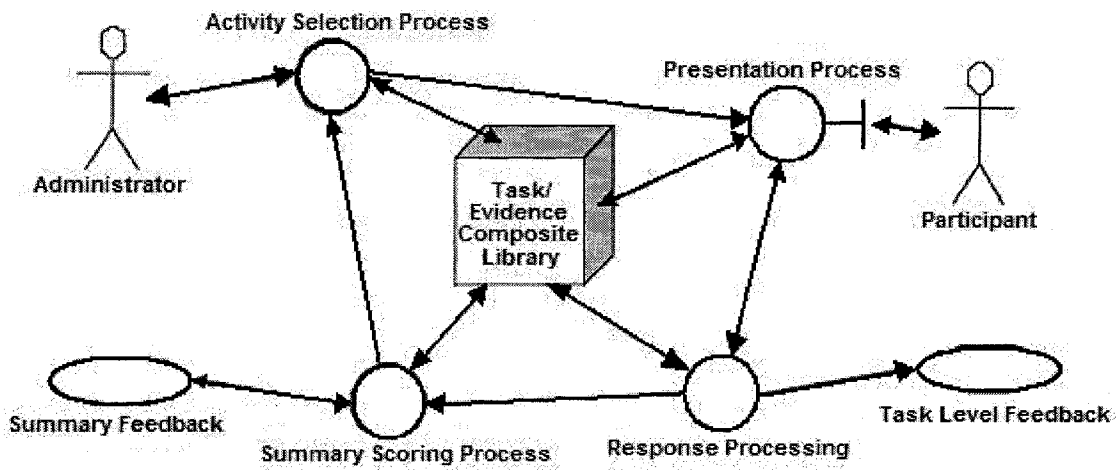

Figure 3. The Four Principle Processes in the Assessment Cycle (Almond, Steinberg, & Mislevy, 2003).

Shepard (2001), in considering instruction and assessment, presents a compelling argument for a way ahead in classroom assessment. Her approach combines cognitive and constructivist learning theories, re-visioning of curricula, and, of course, formative assessment. The goal, she states, is to create an effective and appropriate learning culture.

Key to her model is the concept of dynamic and on-going assessment, though not necessarily testing. Additionally, feedback and movement toward student self-assessment is considered vital in the learning culture sought. On a teacher's level, Marzano (2003) identifies three critical factors to instruction. First, he states that it is necessary to develop many instructional strategies. The interpretation is that the larger the number of available strategies, the more likely it is that one will be found that allows specific students to grasp specific content successfully. Cohen and Ball (1999) referred to this point as a teacher's "repertoire of means" (p. 3) to create appropriate environments in presenting that which is to be learned. Marzano's second point is that instruction cannot occur if the environment is not appropriately ordered. Thus, classroom management that includes rules and procedures as well as addressing the relationship of teachers and learners is necessary. Marzano's final factor is the appropriate sequencing and pacing of instruction in accordance with a plan. In most cases, standards suggest an annual timeframe while district curricula and pacing guides seek to dissect the task into appropriately sequenced and more consumable bits. Of Marzano's three points, planning is critical to points one and three, especially with regard to assessment.

Beyond teachers' abilities to relate to and interact with students, Cohen and Ball (1999) state that instruction depends on a teacher's ideas about learning, their knowledge of and comfort with content, and their flexibility. McMillan (2003) suggests that expert teachers are usually more skilled in their ability to monitor and interpret the complexities of instruction. He further states they are usually more reflective and seek to interpret, evaluate, and explain what they perceive in the classroom and, in doing so, generally demonstrate clearer direction and employ better procedures while being more meta-

cognitive in their effort to understand their classrooms. Finally, he contends that the expert teacher is usually more alert to visual and auditory cues in the classroom and constantly plans for the complexities encountered. While McMillan addresses the capacities of expert teachers, the goal is to have such expertise infiltrate all teachers.

Delandshere and Jones (1999) focus on the content suggesting that teachers must know the content area. While that may sound disparaging, elementary school teachers are normally generalists. In that regard, they may have a preferred content area(s) but in a classic model of elementary school instruction, the teacher prepares for and teaches all content regardless of preference or focus of education. When referring to standards, the standards written by experts in a content area are generally based on a thorough knowledge of a field of study and all that any one standard might entail. A more generally prepared teacher might, potentially, be less capable of seeing the significance unless prompted or trained to do otherwise. Content knowledge should be sufficient to permit feedback, suggests Shepard (2001), that identifies key errors in student thinking, determines the likely reason the error was made, and assists the student in avoiding the errant pattern in the future. In this regard, Cohen and Ball (1999) state that instruction is affected by teachers who continuously seek ways to expand their own content knowledge and develop their own capabilities to affect instruction. The formative assessment and testing concepts would also require additional education or training until such time that necessary capacities are included in existent preparation programs.

Planning in all things educational is critical. Until the recent increased interest in formative assessment and testing, assessment was an element often considered separate from instruction and added after instructional planning and development were complete

(Pellegrino, Baxter, & Glaser, 1999). Glaser and Silver suggest that the two elements were largely "decoupled systems" (1994, p. 403; 1994, p. 14). In a decoupled model, any merit that assessment might have had in contributing to achievement was lost or became obtrusive as it was not well integrated, meaningful, linked, or well-received. Gronlund (1998) contends that to be fully integrated, assessment must be planned with instruction.

When assessment is woven into instruction as a means of obtaining a constant feedback source for further instruction, it blurs the line between instruction and assessment (Anderson & Krathwohl, 2001; Brookhart, 2003; Koretz, Mitchell, Barron, & Keith, 1996). The use of constant feedback to keep a system of instruction stable and properly operating is a concept well understood in most fields with on-going operational processes. The instruction - assessment feedback loop may be aptly represented by the use of formative testing results to guide subsequent instruction. Each step of the process, in and of itself, must be well-founded and able to appropriately interface with other components in order to produce the desired responsiveness and stability. This is consistent with Glaser's Basic Teaching Model (1962) and Almond, Steinbert, and Mislevy's (2003) Four Principle Processes of the Assessment Cycle (i.e., activity selection, presentation, response processing, and scoring, Figure 3; 2003). The failure of any element has the potential to compromise the entire process. A qualified system administrator or operator, in the form of a teacher, is crucial. The components required for successful assessment or testing, especially when used in a formative mode, are also reliant on both human and modern digital components each of which maybe fraught with the potential for failing to deliver the valid appraisal results desired.

Other experts in the field suggest additional considerations in the planning

process. Buckendahl, Impara, and Plake (2002) suggest 1) that instruction and assessment meet state standards, 2) that all students have the opportunity to learn, 3) that instruction and assessment be free of biases, 4) that assessment methods be appropriate for the stage of learner development, and 5) assessments or tests be reliable and consistent. To point one, Baker (2002) adds that clumping standards together is not good for either instruction or assessment as it reduces the clarity of feedback, a point Popham (2003) supports by his suggestion that instructional planning at the substandard level is appropriate. To these planning points, Anderson and Krathwohl (2001) would suggest that a cognitive taxonomy is useful. Such approaches permit teachers to knowingly choose the path students are to take and the cognitive structures they are to build while learning the required content. Haladyna (2004) supports this point suggesting that one must align content and cognitive processes in any instruction and assessment.

A final planning consideration is a proactive synthesis of suggestions from studies examining the results of assessment programs. Koretz, McCaffery, and Hamilton's (2001) evaluation of what can happen when teachers respond to assessment outcomes suggest that there are seven possible forms of effort that result. Always positive actions include: 1) teaching more, 2) working harder, and 3) working more effectively. In planning for instruction, these are appropriate strategies that underlie other considerations such as content, methods, or the direction provided by formative testing results. The next three considerations include 4) reallocation, 5) alignment, and 6) coaching in preparation for assessment. Depending on how they are implemented in instruction, they have the potential to be either productive or damaging to learning. The most important of these in the planning phase is the appropriate allocation of instructional time to curricula aligned

to standards. The plan should call for balance. The seventh point, cheating, is always negative. In the planning process, methods that would disallow cheating are appropriate. A noteworthy addition to the discussion of test preparation through instructional means is addressed by Haladyna, Nolen, and Hass as they labeled nine similar instructional activities as either ethical or unethical (1991, p. 4). Ethical steps include instruction in "testwiseness", test taking strategies, checking answer sheets for completion, and motivating the learner through appeals for assistance from all stakeholders including parents. Unethical steps include basing curricula on tests, basing instruction on test items, using items similar to those on a test during instruction, or the use of score boosting activities. To these the authors add two highly unethical practices including dismissal of low achieving students on test day and the verbatim use of items from the test while preparing students. Generally, controls for these concerns may be effectively incorporated during planning activities.

In summary, instruction and planning for assessment is complex and has been suggested as a departure from past practices. This is not to say that once designed, the plan should be inflexible. Proper planning, if for no other reason than a teacher's consideration of and familiarity with the elements and possibilities, provides the potential for informed flexibility. Instructional planning that lays the ground work for the formative assessment or testing cycle is critical to its success.

Assessments and Formative Testing

When describing the state of classroom assessment, Stiggins (2001) uses the phrase "victim of gross neglect" (p. 10) having suggested that assessment literacy amongst teachers and administrators is "unacceptably low" (p. 5). He explains that the

lack of assessment literacy results in inaccurate assessments, ineffective feedback to students, and a net impediment to achieving learner potential. While Stiggins and others have, for decades, suggested a low state of assessment literacy amongst professional educators, Stiggins is quick to point out that the scope of the problem entails retraining about 2.5 million teachers and administrators (1999), not something subject to a quick fix. As a net result of poor literacy or preparation, McMillan (2003) reports "that teachers' assessment decisions were largely based on on-the-job experience" (p. 38). McMillan considers his observation to be consistent with Cizek, Fitzgerald, and Rachor's findings (1996). In this regard, assessment practices, he states, become highly individualized and seldom conform to "best practices" (p. 34). One might suggest, given McMillan's observations, that the culture or knowledge of assessment and testing within a given school would then be perpetuated in the transfer to and practices of new teachers. If the culture created by more experienced faculty is acceptable or better, then the literacy is likely acceptable or better; if not, the literacy and future of assessment maybe less so.

A relatively complete solution is advocated by McMillan (2003) who states that there are eleven points that must be considered by teachers in their assessment and testing endeavors. Many of these translate into knowledge they must have as a foundation. First, assessment has to be come part of a teacher's beliefs and values. Second, teachers must become aware of the affect that assessment has on students including the impact on their efforts, motivations, and self-image. As a third point, McMillan states that there is a need to be able to individualize assessment. This capacity, he suggests, requires more development. Teachers need help in aligning their classrooms to the high-stakes environment of standards assessment is McMillan's fourth point. While self-assessment

is the ultimate goal for each student and needs to be pursued, time and effort have to be expended to improve what is known of formative assessment as defined by Black and Wiliam (1998a, 1998b) and Stiggins (2001). These constitute points five and six. Pre-service education is the focus of point seven, suggesting that the issue of integrating assessment with instruction has to be required in teacher preparation programs. Recommendation eight cites the need for instruction in measurement, objectives (standards), and taxonomic considerations. As point nine, McMillan suggests that during the education or training of pre- and in-service teachers, they should be afforded the same considerations as the students they teach including their varied circumstances and individual needs. This factor might also ameliorate inculcated negative beliefs about education and teachers. Point ten addresses a culture of learning as being important and assessment being considered a natural part of the process. The final point seems reconciliatory between psychometricians and teachers in that the educators bringing the concepts of measurement to teachers must be more sensitive to teachers' situations and understand their plight rather than be the bearer of psychometric truth.

For those providing education and training in assessment, the realization that testing programs elicit sensitivities must be evident. Lewis (2001) warns that political influences and new technologies allow few implemented testing programs to actually succeed. Marzano (2003) suggests that a first mistake that many schools and districts make is the purchase of off-the-shelf solutions addressing testing needs without proper regard for their content and congruence with what is being taught. Unaligned tests, more often than not, assess curricula very differently from that which is taught and learned by students. These testing efforts only further distort beliefs about testing and alienate

teachers and students alike. Moreover, such tests are insensitive to any real gains achieved. Marzano (2003) contends that the ineffectiveness of these "indirect measures," as reported by Madaus, Kellaghan, Rakow, and King (1979), is a problem as opposed to "direct measures" that are particularly sensitive to the district or school's curricula. Marzano also identifies a second mistake in that testing programs often have no plan or system to address the uses of the data it generates, the analyses to be made, nor the use of findings. To this, Lewis (2005) adds the erroneous belief that technology will solve testing problems by making it possible to process and obtain voluminous amounts of data. Knowing how to employ the data generated is key.

The use of formative testing requires some additional and selected teacher awareness. This includes an elementary understanding of test items, test structure, administration and processing, analysis, and results leading to informed instruction.

Items. Teachers should be familiar with proper test items. The first goal is to examine items for relevancy ensuring that their learners were exposed to the tested materials during prior instruction. Teachers should also be able to decipher item statistics. If items included meet for-learning goals, then insights to learner errors may likely be linked to the examination of the distractors used. Conventional guidance suggests a "plausible" connection of distractors to an item (Popham, 1971, 1999; Gronlund & Linn, 1990; Mislevy et al., 2002; Haladyna, 2004) although Popham (1971) suggests that items must match the purpose of the testing process. In a formative process, responses should challenge the patterns of thinking used by the learner. Distractors in tests for learning should be based on likely procedural errors, faults in logic, errant lines of reasoning or understanding, and anticipate incorrect interpretations. These should be obvious to

teachers. Mathematics is one content area that lends itself to mechanism such as error patterns (Ashlock, 1994). Properly considered, distractors can provide a more diagnostic source of information suggesting why the student did not respond correctly and shorten the path to revised instruction.

Test structure. The attraction of multiple-choice, fixed, or selected response testing has been known since the early twentieth century (Madaus & O'Dwyer, 1999). The expense of multiple-choice testing, in time or dollars, is as much as 60 times less than alternative forms of testing (Stecher et al., 1997). A key point for teachers involved in formative testing processes is that results are usually available more quickly. This category of testing has been, as attested to by Stiggins (2001), McMillan, Myran, and Workman (2002), McMillan (2003), and others as the likely foundation of modern testing, test theory, test measurement, and psychometrics though there is still more to be learned. Properly designed and articulated content, cognitive tasking, difficulty, and a myriad of other aspects can be accommodated by a well-constructed, multiple-choice test. Teachers should be familiar with these concepts and what constitutes clearly correct, best, plausible, or diagnostic response options as addressed in the literature (Popham, 1999, 2004; Cunningham, 1986; Haladyna, 2004; Gronlund, 1998, 2006; Gronlund & Linn, 1990; Stiggins, 2001a; Thorndike, 1997).

In tests used for formative purposes, teachers should be aware of the number of items dedicated to a construct or sub-standard. The psychometric view suggests that it permits the proper estimation of internal reliability. For teachers, it permits redundancy while allowing an examination of different taxonomic and difficulty levels. It should also be apparent to teachers that the difficulty and cognitive tasking of items might also

explain drops or rises in resulting scores as the item's difficulty may be responsible for noted trends and not the degree of learning. The literature suggests that variations in difficulty should be considered critical elements (Berk, 1980; Crocker & Algina, 1986; Cunningham, 1986; Gronlund & Linn, 1990; Cizek, Bunch, & Koons, 2004; Haladyna, 2004). Discrimination indices, while typically used for summative purposes, are also telling statistics for teachers (Berk, 1980; Crocker & Algina, 1986; Cunningham, 1986; Gronlund & Linn, 1990; Cizek, Bunch, & Koons, 2004; Haladyna, 2004). Discrimination indices permit a teacher's quick view of the nature of success or failure in a classroom as items with large positive values suggest a wider range between high achieving and low achieving learners while negative, low, or moderate values are more ambiguous and might suggest a lack of understanding attributable to a whole group or to test item errors. These insights can assist informed instruction.

Administration and processing. In formative testing, teachers should be aware that there are steps in the process that potentially jeopardize the results used to inform instruction. There is, literally, no room for error. An errant report of a learner's achievement can be devastating to the learner if not to the credibility of the testing effort at large and can be the basis of inappropriate revisions to instruction and strategies. As Stiggins (2001a) suggests, these inaccurate assessments and the errant feedback generated have the net potential of impeding learner progress. Sources of possible error under the control of teachers are many. Simply recording, for example, history test answers on a form processed as mathematics will most certainly result in errant findings. Once answers are recorded, any number of mishaps can occur in a less than organized process, each one capable of compromising perceptions of a learner as well as the virtue

of the process. While there are reasons mishaps occur, in a serious effort to inform instruction, any error is fatal to one or more aspects of the process. From a psychometric perspective, these mishaps constitute a source of measurement error that diminish reliability and the potential usefulness of all results in addition to misrepresenting the individual (Crocker & Algina, 1986; Bedwell, 2004; Wong & McGraw, 1999).

The best way to protect against sources of error is proper administration and processing. Teachers normally are a part of these processes. Koretz, Stecher, Klein, McCaffery, and Deibert (1993) suggest that testing conditions be standardized as these measures generally improve the quality of student performance data. Stecher and Klein (1997) also advocate that standardized tasks and procedures make the processes easier to accomplish without extensive training. For the various responsibilities and functions involved, Stecher, Barron, Borko, and Wolf (1997) recommend each task be defined and assigned as well as identifying necessary resources. Their argument is that the clear "allocation of responsibility" (p. 11) must be understood in order for valid testing to work in concert with instruction. The authors also point out the tendency to underestimate the logistics involved and the impact that those logistical requirements have on teachers. Anything that demands more teacher-time is detrimental to the greater purpose of formative testing to inform instruction.

As technology is a part of the formative testing process, teachers must be aware of the pros and cons. In addition to the potential "Garbage In, Garbage Out" syndrome oft associated with digital processing, automated processing creates a mass of data that must be further processed into something teachers can use (Lewis, 2005). The additional processing is normally done by teachers. Ravitz (2002) points out the importance of

technology in today's formative testing efforts stating that while assessment paradigm shifts may be underway, the acquisition of appropriate technology may not be. Moreover, the additional burdens created are heaped atop existent teacher requirements. Time-savings must also be realized through technology. Given that appropriate technology is available to assist in the process, timeliness is critical to a skilled teacher using the data. This may be one of the more critical aspects.

Analysis. Once test responses are converted into useful data, the data must be analyzed. Table 4 lists the data and reports available to teachers within the district studied during school year 2004-2005. These are relatively typical of reports available from five known providers of like service. The amount of data available is overwhelming to the

Table 4
Statistics and Reports Available for Teacher Analyses (Software America, 2005)

Report Series	At-Risk Report Series
alpha rosters	alpha rosters
ranking rosters	ranking rosters
item analysis including response counts	item analysis including response counts
p-values	*p*-values
discrimination indices	discrimination indices
measures of central tendency	measures of central tendency
content validity estimations	content validity estimations
frequency distributions	frequency distributions
alphabetic student responses	alphabetic student responses
ranked student responses	ranked student responses
mastery by category	mastery by category
mastery by standard	mastery by standard
mastery by substandard	mastery by substandard
item analyses by category	item analyses by category
item analyses by standard	item analyses by standard
item analyses by substandard	item analyses by substandard
summary report by category	summary report by category
summary report by standard	summary report by standard
summary report by substandard	summary report by substandard
summary report by success criterion (pass advanced, pass proficient, fail)	summary report by success criterion (pass advanced, pass proficient, fail)

unprepared. Teachers who must make sense of it need what McMillan refers to as "essential measurement evidence skills" (2000, p. 2). Given that these data are error-free, it is, as suggested by Shepard (2000b), the "systematic analysis of evidence" (p. 8) that is crucial. McMillan (2000) advocates that teachers need the ability to understand and conceptually interpret statistics, not necessarily compute them. Computations should be transparent to teachers as it is likely that they do not have the time, inclination, nor understanding to generate them (Black & Wiliam, 1998b; Yoon & Resnick, 1998; McMillan, 2003). In addition to being familiar with the measurement concepts involved, teachers need a modicum of adeptness so as to reduce any delay between testing and revised instruction. The goal, then, is to use the data to make decisions concerning instructional treatments befitting each student.

Brown and Capp (2003) suggest that data-driven instructional decisions must be explicit. They contend that a synergy is created when the data are properly analyzed and reveals what items are of interest to the class, groups within the class, and individuals. The value of this expenditure of time is simple as demonstrated by an example in Stecher, Barron, Kaganoff, and Goodwin's (1998) report on the classroom practices of teachers in Kentucky. They reported that 80 percent of high-gain schools conducted substantive testing in mathematics where only 35 percent did so in low-gain schools. With the assistance of standards-based testing processed and partially analyzed by technology, the only thing that remains is the applications of appropriate strategies to the class, groups, or individuals revealed as requiring assistance.

Results leading to informed instruction. Gronlund (1998) suggests that the process of informing instruction begins with identifying the student's successes and

failures and making instructional adjustments accordingly. Determination of success and failures lie with test results collected and analyzed. Linn (1998) suggests there is increasing interest in differentiated instruction. This is borne out by in the literature suggesting that instruction needs to be specific for each learner (Athanases & Achinstein, 2003; Brimijoin, Marquiesse, & Tomlinson, 2003; Council of Great City Schools, 2003, 2004a, 2004b; Sachs, 2004; Yorke, 2003). Therefore, adjustments to curriculum, instruction, strategies, and methods of approach should be attributable to groups and/or individuals. Gronlund (1998) advocates that prescriptions must be associated with constructs as well as students and that relearning should begin immediately. These views are shared by Black and Wiliam (1998a). The problem lies in whether teachers are prepared to accomplish sub-standard by sub-standard, synergistic analyses, and revisions of instruction by class, group, and/or individual (Brown & Capp, 2003). Given that analysis skills are sufficient, unless the revision of instruction recommended by the data can be planned and implemented, the formative testing program is likely to be a hindrance to learning or simply relegated to being a predictor of future summative assessment outcomes. If, as Gronlund (1998) and Marzano (2003) suggest, that formative testing be done often, teachers must be well practiced in the process. Knowing how to teach more, and how to elicit harder and more effective work clearly become skills that could change future outcomes (Koretz, McCaffery, & Hamilton, 2001).

Summary

If teachers have not been exposed to the issues of beliefs and their power, and rudimentary understanding of standards, pacing, instruction, assessment and formative testing, items, test structures, administrative and processing demands, analysis, and

transformation of results into new instruction, one might have doubts about their preparedness to be in a classroom in an age of standards and assessments. Many of these elements are included in the content of existent coursework, courses, and training, but not all. However, these elements are variable and seldom presented as components of cohesive systems (Shepard et al., 2005). Referring to Almond et al.'s Four Principle Processes of the Assessment Cycle, each of the elements discussed in this section add some insight to each of the principle processes that teachers should know and have the requirement to accomplish. These are elements one would hopefully find in undergraduate and pre-service programs, in graduate education, degree-seeking or continuing, and in training or professional development.

Undergraduate and Pre-Service Education

Undergraduate education and teacher preparation programs are the paths through which most enter the profession of teaching. In the age of standards and assessments, the issue of whether teachers are prepared for standards-based testing environments including formative assessment is a matter of interest. A key issue is whether undergraduate or pre-service education provides alternatives to McMillan's "on-the-job" observation (2003, p. 38). The issues of teacher preparation programs as they pertain to beliefs, content, and assessment, formative and/or summative, will be considered in this section.

Tangible standards for what constitutes a viable teacher undergraduate education or pre-service program are specified by various cooperating organizations. Kim, Andrews, and Carr (2004), in their examination of teacher education programs, refer to the Interstate New Teachers Assessment and Support Consortium's (INTASC) standards as a means of comparison (1992). As reported by Valli and Rennert-Ariev (2002), INTASC

competencies have been embedded in the National Council for the Accreditation of Teacher Education (NCATE) standards. The INTASC principles or standards considered important are listed in Table 5. These are general although reasonably clear concepts. Of the principles listed, content knowledge, learning, instructional strategies, curriculum and planning, and assessment skills have been previously discussed as areas with which teachers should be familiar. Assessment, by comparison to mathematics, is less definitively described.

Table 5
Interstate New Teachers Assessment and Support Consortium (INTASC) Beginning Teacher Standards (INTASC, 1992)

Principle	Teacher's Understanding
1	Knowledge of content areas taught; ways of engaging students in learning
2	Child development and individual ways of learning
3	Diversity of students and variations in their ways of learning
4	Instructional strategies and methods encouraging development of critical thinking, problem solving, and performance skills
5	Engagement of individual/group dynamics, motivation, and behaviors
6	Effective communication in all forms
7	Comprehensive instructional planning
8	Assessment and evaluation strategies
9	Reflective evaluation of self and decisions, students, stakeholders
10	Development of relationships with professional peers and stakeholders for the sake of student support and well-being

Regarding content knowledge in mathematics, NCATE, in conjunction with the National Council of Teachers of Mathematics (NCTM), details program standards for the initial preparation of all teachers of mathematics (1998). For kindergarten through fourth grade (K-4) teachers, program requirements specify 15 semester hours in mathematics and mathematics education addressing specific objectives. The competencies clearly support Virginia's SOLs as both SOLs and education program competencies share a common origin in NCTM's National Standards for Mathematics (1989). Virginia

licensure requirements for early childhood and elementary teachers only differ in the number of semester hours required for licensure (9 versus 12, respectively) which are both less that those required by NCATE. Specific competencies are included from several conceptual areas. Teachers must be knowledgeable of number systems, structures, operations, and properties. These lead to an appreciation of number theory, ratios, proportions, and percentages. The basic concepts of algebra are required to include operations, fractions, equations, inequalities, radicals, exponents, sequences and series, functions, and graphical and tabular transforms and representations. Geometry must also be points of knowledge including properties and relationships of geometric figures, deductive and inductive reasoning skills, concepts of perimeter, area, and volume of 2- and 3-dimensional figures. Probability and statistics are required knowledge to include permutations and combinations, probability, prediction, measures of central tendency, normal distributions, and various plots. A rudimentary understanding of mathematics as it applies to computers, programming, and computer applications is also considered important. Regarding the nature of mathematics, the sequential nature of mathematics as well as its concepts and procedures, necessary reasoning, problem-solving, and effective communication of mathematics ideas are necessary content. From a societal point of view, the licensee must be mindful of the contributions made to mathematics by various cultures, the impact that mathematics has had on cultures and societies, as well as an understanding of the technology that has both influenced and been influenced by it (Virginia State Board of Education, 1998). It is noteworthy that these specific interests are near directly extracted from NCTM Program Standards for Initial Preparation Programs (1998) and equivalent to strands of knowledge from which the SOL are created.

Both assessment (Standard 3) and mathematics (Sub-Standard 4c) are addressed by the joint National Association for the Education of Young Children (NAEYC) and NCATE Standards for Early Childhood Professional Preparation Initial Licensure Programs (2001). The Association for Childhood Education International (ACEI), in conjunction with NCATE, also addresses requirements for elementary education standards for both mathematics (Standard 2.3) and assessment education (Standard 4) in preparatory programs (ACEI, rev 2003). It is clear that national and international teachers education oversight organizations are attempting to attend to the need for specific preparatory exposure for pre-service teachers.

In addition to the standards, principles, and competencies either stated or cited, the literature makes a case for the inclusion of and teaching with an appreciation for the power of beliefs on understanding and use of the learning processes as well as greater emphasis on measurement and formative practices (Borko, 1997; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Prawat, 1992; Black & Wiliam, 1998a; Black & Wiliam, 1998b; Darling-Hammond & Ball, 1998; Delandshere & Jones, 1999; Cohen & Ball, 1999; Shepard, 2000a, b; McMillan, 2001; Stanford, 2001; Gerges, 2001; McMillan, Myran, & Workman, 2002; Hamilton, McCaffery, Stecher, Klein, Robyn, & Bugliari, 2003; McMillan, 2003). One summary of these concepts is offered by Johnston, Guice, Baker, Malone, and Michelson:

> School assessment practices operate at many levels, from the moment-to-moment assessments teachers make in the classroom to the use of standardized tests for institutional decisions and individual placements, and "in the real world" assessment practices at all levels are motivated and

> sustained by systems of beliefs about teaching, learning, and the domain being assessed. Different belief systems produce different representations. The classroom teacher is the point at which all of these layers of assessment and any curricular innovations come into contact (1995, p. 368).

The inclusion of beliefs is important. Beliefs appear to be at least a catalyst in many other aspects of teaching and teacher preparation. Otero (2006) argues that teacher programs are not aligned with new teachers' beliefs. If true, it is an issue of major consequence as Sutton, Cafarelli, Lund, Schudell, and Bichsel (1996) suggest that pre-service beliefs limit the impact of the fundamental ideas incorporated in teacher education. One connection of these points suggests that institutions may not be prepared to provide the instruction that turns the tide of beliefs. These and other researchers make a case for beliefs being appropriate learning goals early in pre-service and undergraduate teacher preparation programs. The targets are varied and include ideas about learning, students, and students' capacities to learn (Black & Wiliam, 1998a, 1998b; Hiebert, 1999; Jones & Vesilind, 1995; Stecher, Barron, Borko, & Wolf, 1997; Taylor, 1994), beliefs as filters of learning (Barlow & Reddish, 2006), beliefs about assessment and testing (Cizek, 1993; Shepard, 1991; Delanshere & Jones, 1999), and the tensions that may occur between teachers' beliefs and external forces such as assessment or testing requirements and the institutions that impose them (Black & Wiliam, 1998a; McMillan, 2003). One simple restatement of these authors is that if one does not believe in formative assessment, it is likely not going to be effective.

Content to be taught is obviously fundamental to undergraduate and pre-service

teacher preparation programs. While the classic, generalist, interdisciplinary, elementary school teacher must share their preparation hours among all content areas, the content may be a larger issue in determining success (Cohen & Ball, 1999; Delandshere & Jones, 1999). In the area of mathematics, courses taken in addition to those that are part of a particular licensure program and methods courses, meeting joint NCATE/NCTM standards, should then be helpful in the teaching of mathematics as would any additional courses in any content area for the teaching of the respective content.

Of seven Virginia colleges of education catalogs or program bulletins reviewed, all required at least one mathematics instructional methods and strategies course, while mathematics courses, in addition to general education requirements, ranged from one to three. This is likely in response to the Virginia Licensure Regulations. An effect of additional content courses is reported by Childs, Ross, and Jaciw (2002) stating that students of teachers with mathematics degrees had higher test scores in mathematics than students of teachers with undergraduate degrees in education. Such a foreboding overtone suggests a matter for further exploration.

If preparation to teach given content is a substantial variable, the aspects of assessment may be even more uncertain. Few teachers, Stiggins (2002) contends, are properly prepared for any classroom assessment. Assessment, formative or otherwise, is an important topic in pre-service or preparatory teacher programs. Zemelman, Daniels, and Hyde (1998) suggest "effective teachers are aware of the importance of a thoughtful, systematic approach to assessment" (p. 24). Their tendency to value formative techniques that aid in understanding a child's development assists them in making informed instructional decisions. For those that do offer instruction in assessment, Otero (2006)

argues, that it is a practice taught often disconnected from learning theory or concepts. The distinction between summative and formative assessment, for example, is a nuance that Stiggins (1991) suggests is vital. As new teachers discover, there is a perpetual need to produce grades for report cards and reports for parents. The enticement to grade is obvious. The advent of high-stakes testing makes such an orientation even more compelling. The concept of assessment for learning is lower on the agenda. To make matters more tenuous, McMillan (2003) suggests that teachers can seldom provide reason for their assessment and grading practices. McMillan summarizes research on the topic as revealing "a highly individualized, idiosyncratic process, one that did not seem to be founded on common assessment principles" (p. 38).

Regarding assessment instruction in teacher preparation programs, the literature seems to have a negative overtone. Cizek (1998), for example, reports that the widespread lack of fundamentals, principles, and practices of assessment, citing O'Sullivan and Chalnick (1991) and Ward (1980), is real and will not change until there is a prerequisite for licensure. The argument for assessment standards suggests that if there is no compelling reason to include them as licensure requirements, teacher preparation institutions are not likely to respond. This is in light of professional teachers and teaching oversight organizations (i.e., NCATE, NCTM, NAEYC, INTASC, and ACEI) that advocate them. Examples of advocacy include INTASC's Principle 8 concerning assessment (see Table 5) and NCATE's approval of the National Association for the Education of Young Children's (NAEYC) Standards for Early Childhood Professional Preparation (2001), both of which clearly specify assessment as a standard component of initial licensure programs to include the need for understanding of

distinctions and definitions as well as various psychometric concepts such as validity, and NAEYC's less specific standards for graduate programs (2002). It is also apparent that the issue is not viewed as particularly relevant as only 12 of 50 states specify such requirements (Stiggins, 2002). The arguments against assessment standard requirements are not as visible. The rebuttal is often formed in the simple accounting of program hours; what existent requirement is to be dropped in lieu of assessment? In the absence of assessment education, the potential for misuse of assessment or development of formative assessment skills, Stiggins (2002) contends, is a day-to-day concern.

As a component of assessment, knowledge of measurement employed in the formative testing of learners is necessary. The examination of instruction in measurement in teacher preparation programs reveals observations such as Stiggins and Bridgeford's (1985b) which states "many do not require measurement training and teachers often avoid it, given a choice" (p. 284). Such an attitude states the gravity of the problem. Citing Cizek, Fitzgerald, and Rachor (1996), McMillan (2003) states that assessment decisions are "largely the result of on-the-job experience" (p. 38). As such, he continues, there is little derivation from scientific measurement or current theory. In a balanced counter-argument, he further suggests that some commonly used measurements may be irrelevant to a teacher's day-to-day needs. McMillan (2000) clearly states that teachers need the conceptual knowledge of descriptive statistics. Crocker (2003) advocates that more knowledge in the area of measurement and research is required thereby preparing teachers for ever changing roles in instruction and assessment. Anecdotally, it is interesting to note that in the introduction to several high quality texts on assessment, measurement, and testing, the authors (Anderson & Krathwohl, 2001; Cunningham,

1986; Gronlund & Linn, 1990; Haladyna, 2004) suggest their respective works are worthy of inclusion in undergraduate teacher preparation or graduate programs. Once measurement skills produce information worthy of action, the next concern is determination of what that action might be.

McMillan (2003) writes that "it is well established that reflective decision-making is necessary for effective teaching" (p. 35) citing Clark and Peterson (1986), Good and Brophy (2000), and Wilen, Ishler, Hutchinson, and Kindsvatter (2004). This is consistent with NCATE and INTASC standards. The capacity to be reflective is often an issue of time management and a setting of priorities. This is possibly why teachers who treat instructional decisions in more superficial ways make less insightful decisions. McMillan (2003) also asserts that reflective decision-making has theoretical implications, that there are essential skills that should be a part of teacher preparation. Answers to simple questions such as "Was the test trustworthy?" (reliable), "Did we cover that material?" (validity), and "What are these results telling me?" (inference which relates to validity) are, at a minimum, required knowledge in making good decisions. He further suggests that how learning is evaluated, how positive feedback is formulated and delivered to students, and how instruction is revised or informed are equally important lessons. The absence of these objectives, and others, lead him to remark that it "may not be surprising given the well documented lack of appropriate training in classroom assessment in teacher preparation programs" (p. 36).

The capacity to provide instructive feedback is a key skill in formative assessment. Schafer (1993) attributes to Stiggins (1991) and Schafer (1991) the belief that feedback from assessment is a critical part of teacher education. The importance of assessment

feedback is necessary whether it results from a test or, as Koretz, Stecher, Klein, McCaffery, and Deibert (1993) suggest, the evaluation of a portfolio. Costa (1993) suggests that the nature of feedback be non-judgmental.

Finally, Cohen and Ball (1999) suggest that professional norms are the basis for classroom practices that "are strong on individualism and weak on content, common expectations, and standards" (p. 11). They speculate that pre-service education could, but does not, create the foundation to minimize these problems. Teaching candidates need to be fortified against the errant professional norms they might encounter and the impact of external tests and their influence on instruction and assessments. McMillan (2003) reports a sense of resignation on the part of teachers as they modify assessments to conform to external, high-stakes testing demands. This observation has been made in every evaluation of high-stakes testing's impact on the classroom. What is ironic is that the same studies report a contrasting disregard for local, standards-based, district testing.

Graduate and Continuing Education

Graduate education is purposeful in empowering graduate students to master disciplines and conceptual frameworks as well as develop research skills (Heathcott, 2005). Differentiation between forms of master's programs, however, must be made. As few, if any, colleges of education can currently meet all requirements within the confines of a bachelor's degree program, nominally 120 semester hours, many have turned to a fifth year master's degree approach to fulfill all academic requirements for initial licensure. The differences between undergraduate preparation programs and associated graduate studies leading to initial licensure are usually focused on the completion of requirements rather than a classic mastery of a field of study as suggested by Heathcott.

With classic master's degrees, as Mislevy (1996) suggests, graduate students are expected to have sound foundations in their fields. Master's programs other than those leading to initial licensure typically specialize in advanced content and research and evaluation component courses, suggests Eisenberg (1999), and typically emphasize theory, methods applicable to the teaching of a given content area, and statistics. Doheny (2002) suggests that these forms of graduate studies facilitate an examination or re-examination of current research, theory, and practices that are applicable to everyday teaching.

Research is a component typically considered important. For the master of a content field, it is through research that new insights are achieved that can enhance one's teaching capacity. Doheny (2002) suggests that graduate teacher programs are about promoting change in teaching and learning though these types of courses are often challenges for teacher educators. In education, a teacher's insights can generally be realized in enhanced student learning (Esposito & Smith, 2006, citing Burnaford & Hobson, 1995; Johnson & Button, 2000; Sax & Fisher, 2001). Action research is a form considered appropriate by many as it allows teachers to investigate their own practice of teaching, the choices that they make, and the impact those changes have on their students (Esposito & Smith, 2006).

Regarding content, graduate courses often offer a broader range of topics, each more focused on details than the introductory forms experienced in undergraduate settings. An anecdotal examination of catalogs for state universities in Virginia indicate courses in research design, evaluation, instruction, curriculum planning or development, assessment, instructional methods, differentiated instruction, statistics, measurement, and

testing being available. Some programs require courses such as research design, evaluation, and/or statistics while others do not. A necessary component of any good research or evaluation perspective, statistics courses that support evaluation and assessment are also offered. Thorndike (1997), however, suggests that staying abreast of the developments in the field of education is quite different than receiving proper training in a graduate program. The concerns regarding graduate education are possibly carried-over from teacher preparation but are not readily noted in the literature. The combination of pre-service teacher preparation programs and graduate courses are hypothesized as contributory to a teacher's perspective and capacities of interest to this study. These include the influences of advanced curriculum design, learning theory, assessment, testing, differentiation of instruction, statistics, evaluation, and research courses on the teachers capacity to use the formative testing to informed instruction process and ultimately influence student achievement.

There are downsides to graduate programs, however. Heiss (1968) suggests that graduate programs that are not well articulated may find themselves victims of external pressures. From many potential graduate students' point of view, some believe that graduate education teaches in a spiral (Gilbert & Smith, 2003), a never finished process that is frustrating, time consuming, and increasingly more expensive. Those topics not addressed in pre-service or graduate level education are potentially many. Moreover, those that are addressed may be too conceptual for ready use. Topics not covered or not sufficiently focused are the domain of professional development and training. There are also the collaborative efforts of school districts and colleges of education to identify and provide specific courses intended to address recognized deficiencies. The professional

development school concept appears to be one form of this concept (Cochran & Smith, 2003; Darling-Hammond, & Bransford, 2005)

Professional Development and In-service Training

Thorndike (1997) suggests that while graduate education is important, it does not keep one current with new developments in the field of assessment. His contention is that to stay current with changes in theories, practices, and conceptual foundations, one requires constant retraining and professional growth. Obsolete or outdated practices are, as he frames them, a matter of ethics. It could then be said that in-service professional development or seminars constitute the answer.

Professional development, as a term, can mean many things. It can reference the life-long improvement of an individual in any number of dimensions. It can connote the development of a capacity from the contributions of many venues. It can even be attributed to alternatives to classic education paths when used in the term professional development schools. In the case of this study, the use of professional development is limited to the improvement of a teacher's capacity to use testing to inform instruction received from the employing school district, at district, school, and collegial levels. These types of training are often considered in-service. Stiggins and Bridgeford's assessment is that ". . . in-service training, structured to meet teacher's assessment needs, provides the greatest opportunity for impact" (1985, p. 285). In-service training is a reliable, systematic, focused way in which to improve the existent teaching corps of any school. Given that its programs have clear objectives, sufficient resources including time, qualified and exceptional instructors, and a basis in literature and action research, any number of deficiencies in knowledge and skill, preparation, or change in foundations can

be accommodated.

If one is able to get beyond teachers' beliefs about in-service professional development, which Borko et al. (1997) suggest causes participants to either ignore or inappropriately assimilate the training provided, in-service professional development is part of the solution of professional growth. If Stiggins' (1999) estimate that 2.5 million teachers and administrators are in need of assessment training, in-service professional development is probably the path of least resistance. It is Darling-Hammond's (1994a) belief that professional development is important in supporting teachers' needs to embed assessment into their teaching and their students' learning.

The question of whether teachers will accept the training provided is a matter for some concern. Based on a five year study of Kentucky reform efforts in the classroom, Stecher, Barron, Borko, and Wolf (1997) suggest that teachers initially want to understand what the training entails. Teachers want to know the rules, procedures, and guidelines and, like any learner, the objectives and expectations. Teachers will then consider the impact on their teaching. Some will look for greater detail and opportunities. Given time constraints, some will look for efficiency and shortcuts. Some teachers prefer to be proactive, involved in making things the way they should be done while others prefer to be more reactive, being observed, and awaiting recommendations for change. The window of training opportunity is also hard to find as, the authors continue, teachers neither want to attend training during the day nor on their personal time. Their concern is whether the improvements that might be evident justify the expenditure of time. Teachers' perceptions that requirements were 'add-ons' was reported as a source of teacher frustration unless the teacher had already accepted the congruency of the changes

being made to their own goals.

Content of in-service professional development may be a key to teacher acceptance. In-service training in assessment may be a topic that elicits interest as the high-stakes testing environment seems to dominate much of the school year. Such training may be perceived as fundamental to success. Based on Stecher et al. (1997) observations in Kentucky, the value of the content to the teacher may be a path to proactive participation.

Considering commentary concerning teachers' content knowledge, content enhancement would seem apt topics. Stotsky (2006) suggests that most of the money being invested in professional development today is aimed at improving teachers' knowledge of the subjects they teach. In the process of learning more about a content area, it is also prudent to do so in a fashion that blends the desired perspective on formative assessment seamlessly. This requires planning on the part of the instructor and a policy on behalf of the school or district.

Formative assessment in professional development is more than using test results to inform instruction. Cohen and Ball (1999) report that coordinated efforts are required to achieve the assessment training teachers need. The effort is more encompassing than just assessment as a thorough understanding of standards and measurement. Understanding the relationship of curriculum and assessment is also required. Cohen and Ball's (1999) statement that assessment is often developed with little sense of the curriculum is important. The authors contend that while instruction in these matters is profuse, it is often inconsistent and that:

There is a greater volume of guidance in the United States than in other

> nations: teachers and students are deluged with assessments, programs, policies, judicial decisions, instructional materials, advice from pressure groups, and much more. But the guidance is often inconsistent and unclear, in part because the volume of diverse advice overloads cognitive capabilities and encourages superficial acquaintance and misconceptions (p. 11).

Cohen and Ball (1999) contend that, despite training efforts, professional norms seem to drive the classroom and that "professional norms are strong on individualism and weak on content, common expectations, and standards" (p. 11). This strong individualism suggests the existence of separate styles within a school and each of its classrooms. McMillian (2003) concurs referring to these as highly idiosyncratic processes. Cohen and Ball (1999) go on to suggest that, despite the money spent on in-service training, they are "intellectually superficial, disconnected from deep issues of curriculum and learning, fragmented, and non-cumulative" (p. 12). The sense of being "updated" is prevalent to more in-depth training and education.

The specific content of training is a matter for concern. Gullickson (1986) reports that teachers identified a need to understand test preparation, proper administration and scoring, test selection and usage, integration of results with other assessments, appropriate statistics and their interpretation, and use of test results for both formative purposes in instructional planning as well as conventional summative uses. The teachers' prescription seems consistent with many researchers. The effects of desired training can be complex. For example, Cizek (2000), Cromey (2000), Stiggins (2002), and Heritage et al. (2005) state that there seems to be little or no preparation for the use of data in formative ways but Anderson and Postl (2001), Khanna et al., (1999), Ruberstein and

Wodatch (2000), and Ward (1998) report that research of in-service training in the analyses and uses of data suggest a positive influence on student performance.

Formative evaluation, Stiggins and Bridgeford (1985) contend, is also effective in eliminating a teacher's weaknesses. Ebel and Frisbee (1986) expand upon the topic suggesting that a subordinate reason for analyzing tests is to examine a teacher's testing weaknesses to include flawed knowledge or skills and judgment errors regarding items. Such observations prescribe meaningful professional development to improve most aspects of testing and use of the data that it generates.

In an effort to correct many of these observations, Stiggins (2002) recommends long-term professional development programs that create "literacy in classroom assessment" (p. 765), have appropriate resource allocations, support development of large-scale and classroom assessment programs, and require standards for teacher and administrator licensing. These views are consistent with Borko et al.'s (1997) view of staff development programs having recommended that programs be at least a year in length. She suggests that both beliefs and practices must be addressed and – in the resulting, properly reformed environment – learning communities of teachers evolve that consider problem-solving, communications, and conceptual as well as practical understanding of assessment.

Research in the area of professional development has revealed some other interesting aspects. Stecher, Barron, Borko, and Wolf (1997), for example, report that many schools with exemplary learner achievement seemed to have more teachers who were "themselves trainers, cluster leaders, or in-service developers" (p. 20). This suggests that a school or district might well have the preferred instructors available to lead

instruction. This observation is supported by Lewis's report that some districts assign teachers to mentor, train, or develop curriculum based on analysis of test results (2001). Yet another aspect supporting the teacher and professional development relationship is the interest that some researchers have in student performance as related to their teachers' professional development (Heritage, Lee, Chen, & Latorre, 2005).

Given the availability of in-service professional development, correct content, and appropriate instructors, time stands out as a problem. The difficulty in school year training is finding the time for it. Teachers' problems often can be defined as a time-management issue. Black & Wiliam (1998b) summarize the problem as fragile and busy, precluding all but a few from translating principle into practice. Thus, extensive, time-consuming training programs may not be the answer. What the authors recommend is that actions by a small number of formative assessment based and collaborative schools be encouraged and harvested for the benefit of providing the classroom solutions that most other teachers need. This type of action research could pilot the eventual training. The drawback is, of course, time.

As indicated earlier, in-service professional development in assessment has a variety of problems. Impara, Plake and Fager (1993) make the case that where teachers get their assessment skills may be contentious, but the reality is that too many have no formal training in a relatively rapidly changing field and that constitutes a reason for concern. Borko (2004) suggests that much of the professional development available today, despite the money spent on it, is "woefully inadequate" (p. 3). Citing Ball and Cohen (1999) and Putnam and Borko (1997), she estimates the training provided to be "fragmented, intellectually superficial, and do not take into account what we know about

how teachers learn" (p. 3). Bagnel et al. (2006) citing Reis and Westberg (1994) suggests that while educators' knowledge increases as a result of in-service training, minimal change in classroom strategies are often noted. This point may be a matter of perspective or definition of the argument as Cady, Meier, and Lubinski (2006) report that teachers consider professional development to influence change in their teaching. Regarding consideration of the training provided, Anderson and Olsen (2006) suggest that training sessions rarely consider teaching experiences or teacher education programs. Moreover, university teacher educators are seldom involved in ongoing development programs. Several researchers have reported that teachers should be treated as we would have them treat their students (McMillan, 2003).

Until the many views of professional development or in-service programs are better organized, coalesced, and understood, assessment, psychometrics, formative testing, data analysis, decision-making, curriculum redirection, feedback techniques, and encouraging learner involvement are some viable topics for training. The lists of topics in the area of formative testing to inform instruction are approximately the same for undergraduate, pre-service, graduate education, master's degree programs, or continuing education, although more specific instruction in methods, strategies, and instruction probably occurs during in-service sessions. The only differences may be the foci, venues, and instructors.

Summary

In this review of related literature, several contributing topics have been examined. Each has an impact on this research. Standards and assessments were presented as the general field of interest with their potential to constitute a viable educational reform

movement. The topics and differences between assessment of learning and assessment for learning (Assessment Reform Group, 1999) were explored with attention paid to critical aspects of the latter.

The reasoning of things that teachers should know in order to use formative testing to inform instruction led the exploration of various contributing components. Beliefs were addressed as they have the potential to skew many aspects of the process including the acquisition of the capacities to effectively use testing in a formative manner. As a basis for the formative testing cycle, the concepts of state level standards, expressed herein as Virginia's Standards of Learning, and their vulnerabilities were discussed. The dangers of and counters to interpretation were revealed as well as recommendations to prevent the same standard meaning different things to different agents in the process. The concept of cognitive tasking and its importance to standards was also explored. Instructional pacing and its importance and the linkage of assessment and formative testing to instruction were examined, stressing the importance of standards and formative assessment contributions during planning phases. Some discussion was afforded assessment and formative testing and an estimation of the impact that they may have on educators at large and what it might take to correct suggested problems. Some additional topics were briefly elaborated on to include the discussion of items used in testing, test construction, assessment administration and processing, results analysis, and the process of using test results to inform instruction.

Consideration of undergraduate and pre-service education was discussed. Standards for content of said programs were examined, as well as the impact that additional emphasis on content preparation – such as additional mathematics courses –

can have, and the requirements of such programs pertaining to assessment and testing. Teachers' decision-making training and the importance of feedback were touched upon. Graduate and continuing education as contributors to the needed awareness and capacities were also briefly explored. The reality that graduate programs usually suggest content mastery was entered into the equation and that graduate education content was usually much broader and more relevant to desired capacities. The final section of the chapter dealt with aspects of professional development and in-service training as they pertain to the subject of assessment and formative testing to inform instruction.

Given that a teacher understands and accepts the principles that the standards and assessments movement suggests, is versed in formative testing and the components on which it depends, has the capacity to make use of properly detailed test results to inform instruction, and teaches in a school that supports these concepts, then student achievement should reflect the sum of the teacher's readiness. The contributions made by preparatory education, continuing education, professional development, and capabilities resident in teachers and reflective of their schools should be evident. While testing data are relatively easy to acquire, the data that represent these teacher traits are not. Its acquisition is more qualitative in nature. The sum of this research will hopefully shed light on the contribution that each of the educational or training sources has on teachers' capacities to use formative testing properly as defined by their knowledge of standards, planning, instruction, assessment, analysis, feedback, and instructional refinement. Thus, the methods and procedures must be of a mixed design and capable of obtaining the data where it exists, part in school databases and part from the schools and teachers themselves. In Chapter III, these aspects will be detailed.

CHAPTER III

METHODS AND PROCEDURES

The goal of this study was to identify factors in teachers' preparatory education, post-graduate or continuing education, and/or professional development that might be associated with their capacity to use data from formative testing to inform subsequent instruction and thereby improve student academic achievement. The formative process includes the analysis of test results, interpretations of those findings to the conceptual construct or sub-Standard of Learning (SOL) level, and revision of planned instruction to accommodate noted weaknesses. The planning that results can be differentiated by student or groups of students and used for specific improvements in learning.

The steps for using formative testing to inform instruction are cyclic. The teacher's knowledge and skills for accomplishing these tasks are important as they are the basis for the data collected. The data collected identifies the focus of each subject's educational programs, his/her coursework, courses, or training contributing to general analytic and instructional planning skills, and any additional education, professional development, or training that would affect the teaching of third grade mathematics.

It is reasonable to suggest that the sum of a teacher's education and training experience contributes to her/his ability to perform the associated tasks. In an attempt to reveal the contributing factors, this research uses the histories of education and training of individual teachers in select subjects including mathematics, and data suggested by the literature that potentially influence the success of using formative testing to inform instruction with the goal of identifying factors that suggest understanding of the process and the preparation required to accomplish it successfully.

This chapter will first address the type of research being conducted, the population from which the subjects were obtained, and description of the research variables. In this chapter, the design and content of the instruments used are elaborated upon, the methods of data collection explained, the field procedures used to collect data specified, the analyses applied identified, and the criterion considered when making analytic decisions briefly discussed. The chapter is summarized.

Type of Research

This study is an exploratory, *ex-post facto* examination of third grade teachers' education and training histories as they pertain to their respective capacities to formatively transform testing results from quarterly, district-wide, content assessments in mathematics into instruction that addresses students' content weaknesses. The study uses a mixture of quantitative and qualitative data. Students' test results in third grade mathematics during school year 2004 – 2005 were used to stratify teachers into three groups representing students whose performance was characterized as high performance, median performance, or low performance. Teachers' education and training histories represent the qualitative data collected. The research attempted to identify factors or components that represent capacities to formatively use assessment data. Principal component analysis was used to indicate the presence of contributory factors.

Population

The population from which the sample of teacher participants in this study was drawn is comprised of 170 third grade teachers assigned to 35 elementary schools. For the purposes of controlling a number of educational and training variations in these teachers' education and training experiences and a number of organizational variations in

schools, only schools using a classic model of instruction (i.e., all core subjects taught by the same teacher) and employing school-wide testing only when directed by the district were considered. Other instructional models, such as departmentalized assignments in which a teacher teaches one or two core subjects as opposed to four core subjects, change the organization, dynamics, processes, skill sets, and academic goals and achievements of teachers assigned to teach fewer subjects. Schools using school-wide testing on a basis more frequent than that required by the district are likely to have either become more efficient in the process of using testing to inform instruction or have affected the outcomes in ways that confound the meaning of teacher outcomes, such as analysis teams.

In order to obtain a stratified sample of teachers, the Colorado Student Assessment Program (CSAP) Model (Linn & Haug, 2002) was used to rank elementary schools based on 2004 – 2005 district generated content assessments of student achievement in third grade mathematics. These assessments are intended to be formative and inform respective teachers of the strengths and weaknesses of their students' achievement in mathematics thus allowing them to modify instruction accordingly. By selecting the five top, five median, and five low performing schools, it was suggested that contrasts in teacher education and training profiles might become more evident when conducting preliminary analysis in an effort to understand eventual findings. Schools that opted not to participate, were replaced with those that remained in a given strata until no schools remained. There were 71 subjects possible in the stratified sample of 15 schools.

Research Variables

The principle research variables of this study were derived from participation in selected undergraduate or teacher preparation coursework, participation in selected

graduate, master's degree programs, or continuing education courses, and participation in selected professional development or in-service training. While all variables are treated as dependent variables in factor analysis (Field, 2005), the measurement values and the component scores generated constitute independent variables. Instructional variables, based on the various levels of education and professional development and in-service training, are listed in Table 6. Data for educational content as topics or coursework in more introductory courses and content as courses in graduate studies variables were recorded as both participation (yes/no) and the number of class hours or course hours, respectively. For professional development and training topic variables, participation was recorded as both participation (yes/no) while the specific number of hours attended was recorded as an ordinal value, i.e., 0 to 2 hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours.

Nominal scale variables included bachelor's degree major, bachelor's degree minor(s), master's degree major, additional graduate level focus area(s), and courses taken beyond master's degree. Graduate program participation was recorded as a dichotomous variable. Service as an in-service or professional development trainer was recorded in ordinal terms identical to professional development and training received. Responses to questions concerning teachers' personal like or dislike (i.e., preference) for mathematics, belief in formative testing to inform instruction as practiced within the district studied, and perceptions of assistance received from the school's mathematics specialist were recorded using a five-point Likert-scale. The number of years teaching third grade, the total number of years teaching, and class size were also recorded.

Dependent variables associated with this study are the third grade mathematics

Table 6
Research Variables Derived from Teachers' Educational and Training Histories

Education		Training
Undergraduate or Pre-Service	Master's Degree, Graduate or Continuing	Professional Development
	Participated in graduate courses	
Bachelor's degree major	Master's degree major	
Bachelor's degree minor(s)	Additional focus area(s)	
Coursework in:	Courses in:	Topics in:
Assessment Testing Item writing Data-driven instruction	Evaluation Statistics Testing Test item construction Differentiated instruction Research design Assessment Standards Analysis of test data Advanced curriculum design	Mathematics methods Assessment Testing Test results analysis Data-driven decision-making Action research Evaluation Differentiated instruction Strategies, activities, and exercises Instructional planning and revision
Mathematics courses beyond program requirements	Mathematics course in addition to program requirements	
	Courses beyond Master's degree	Service as professional or capacity development or in-service training instructor
	Date Master's degree awarded	

Teaching Experience	
Years teaching third grade	Total years teaching

Opinion or Evaluations (Likert-scaled)		
Personal like/dislike for mathematics	Belief in formative assessment as practiced within the district	Assistance received from school's mathematics specialist

test scores achieved by the students of the subject teachers. These were used to stratify school and teacher groupings for preliminary analyses.

Instrument Design

The primary device used to collect data for this study was based on the 1999 Washington State Survey of Teachers used by Stecher, Barron, Chun, and Ross (2000) in their examination of the Washington Assessment of Student Learning (WASL) Program. Like the examination of WASL, this study examines teachers, their experiences in

preparation and training for teaching, and an assessment system intended to measure achievement. While WASL is summative, similar to SOL Assessments, the authors' report of findings examined aspects of summative assessment that could be converted to examination of formative processes. The original questionnaire was 16 pages in length and addressed multiple content areas and experiences. The questionnaire was modified, with permission of Stecher (personal communication, July, 2005) for this study. The first modification resulted in the reduction of focus to one content area, mathematics. The second was based on pre-project proposal planning discussions with the district's research coordinator. Guidance received stated that all questions of interest were to take no longer than 45 minutes to answer. Some additional modifications were made based on the literature and local professional development training offerings. The resulting protocol was used to guide interviews, survey subjects unavailable for interview, and record review. The protocol is attached as Appendix A.

The protocol was comprised of two pages in seven sections. Section 1 used eight items to query undergraduate or pre-service preparation programs. Section 2 addressed teaching experience. These items asked for the number of years teaching third grade and total number of years teaching. Section 3 addressed graduate level education. Of the 16 items, eleven addressed graduate courses taken involving topics of interest. The remainder of Section 3 dealt with nominal data regarding graduate education participation and the date, if any, that a master's degree was awarded. Section 4 addressed courses taken beyond the master's degree. Section 5 addressed professional development, in-service, or capacity development training attended. Training topics listed were derived from district training catalogs and the mathematics curriculum coordinator.

Section 6 was a single item. Experience as a professional development or in-service trainer often connotes professional interest or expertise. The question was asked in the form of participation (yes/no) and number of training hours provided. Section 7 addressed two items using Likert-scale responses. The first item inquired into the beliefs the subject had for formative testing as practiced by the district while the second asked for an estimation of assistance received from the school's mathematics specialist.

The protocol used to collect information from the district's local assessment and professional development coordinators regarding elementary school instructional and testing models was a simple listing of schools with instructional model and testing model columns. See Appendix B for a copy of the model's protocol.

The protocol used to collect data from the district's mathematics curriculum coordinator regarding evaluations of teachers' capacities to guide instruction based on test results was a simple listing of participating teachers. The protocol for the collection of curriculum coordinator's evaluation of mathematics specialists was a simple list of participating schools. A five-point Likert-scale was used to code these data. See Appendices C and D, respectively, for copies of the curriculum coordinator's evaluation protocols.

Methods of Data Collection

Quantitative data regarding class size, classroom groupings, and other demographic information were collected from the district's student information system. Quantitative data regarding student scores and responses were gathered from an Internet application known as Assessor, a product of Software America. Qualitative data regarding teachers' education and training experiences were primarily collected from

interviews using the protocol found in Appendix A. This same instrument was sent as a survey to teachers in participating schools who did not respond to requests for interview. These data were combined with teacher education and training data extracted from available license certification/recertification records with the permission of respective participants. Both forms of data were self-reported as the content of licensure/re-licensure submissions were left to the discretion of the respective teacher. Qualitative data representing the district's mathematics curriculum coordinator's assessment of teachers' abilities to use testing data to inform instruction were also collected via personal interview.

Field Procedures

Collection of data commenced upon completion of the University's Human Subjects and the district's research project review processes. Schools identified by instructional and testing model, with the assistance of the local assessment and professional development coordinators, were initially contacted by the district's research and evaluation coordinator inviting each principal to participate. Upon the principal's agreement and identification of a point of contact, the researcher was granted permission to contact the designated school representative. After a time for interviews was agreed upon, interviews were conducted within the facilities of each participant school.

The interview format typically consisted of group sessions in which the research was explained, participation was solicited, and informed consent was explained and consent forms completed. For those who agreed to participate, data were collected immediately in compliance with discrete interview time limitation of 45 minutes. The protocol was provided; each item was read and explained, if necessary; and each

subject's response was recorded. Completed protocols, identified only by a random subject identification number, were collected.

For individuals contacted by mail, a brief letter of introduction and explanation was provided as well as all informed consent materials. The protocol, with points of explanation based on experiences with prior interviews, was provided. Pre-addressed and postage-paid mailing materials were included.

Once all interviews and survey data were collected and informed consent forms were in-hand, access to records used for licensure/re-licensure submissions was requested and subsequently made available by the district's human resources department on a single session basis. The interview/survey protocol was used as a format for review of available records. These data were directly transcribed into a digital database.

The final data collection event included brief interviews with the district mathematics curriculum coordinator and district mathematics teacher specialist regarding participating teachers' capacities to perform the conversion of testing data into subsequent instruction and the evaluation of each participating school's mathematics specialist.

Statistical Analyses

Data analyses began with examination of responses for completeness, missing data, and coding errors. All data were transferred from paper protocols to a computer application. Data coded with both dichotomous and interval or ordinal scaled information were identified so as to preclude both forms being used at the same time. Interval responses with a discrete number of hours listed were preferred. If, for any analytic reason, these responses were rejected during the analytic process, the information was

downgraded to the dichotomous scale and calculations re-run. These two data forms did provide a rudimentary opportunity to test instrument reliability. In the case of missing data for education and training, an average number of hours for those who responded with yes within each strata were used. If these items were not answered or a no response was registered, data were coded as no.

Preliminary to factor analysis, a correlation screening was accomplished. According to Field (2005), there is little value in conducting factor analysis with variables lacking sufficient correlation. For initial screening, variables with correlations between absolute 0.5 and absolute 0.9 qualified. Correlation values greater than 0.9 cause potential multicollinearity problems in calculations (Field, 2005; Meyers, Gamst, & Guarino, 2006). Consequently, these data were not used. Descriptive statistics and independent *t* tests were conducted in order to determine strata differences.

Qualifying variables were then subjected to principal component analysis in an effort to reduce the variables in number and potentially indicating the existence of underlying constructs for further examination. Stevens (2002) justifies principal component analysis as being psychometrically sound, mathematically simpler than factor analysis, and a way to avoid some troublesome features inherent to factor analysis.

Based on a review of factor analysis literature and use of SPSS as an analytical tool, criteria considered important in the conduct of the analysis included:

1. Communalities in principal component analyses with a mean less than 0.7 were considered suspect (Field, 2005; Stevens, 2002).

2. Kaiser-Myers-Olkin statistics less than 0.5 were considered inadequate, 0.5 to 0.7 adequate, 0.7 to 0.8 good, 0.8 to 0.9 excellent, and greater than 0.9 superior (Field,

2005; Meyers, Gamst, & Guarino, 2006).

3. Bartlett's test of sphericity should be significant, $p < 0.05$ (Field, 2005).

4. Sample size considerations for factor analysis or principal component analyses were based on the research of Guadagnoli and Velicer (1988). Their findings suggest that factors with four loadings of greater than 0.6 are valid regardless of sample size and three factor loadings greater than 0.8 are also valid (Stevens, 2002). General guidance suggested that any factor loading over 0.6, when considering small *n*-size, were worthy of retention for re-examination with larger sample sizes.

5. Kaiser's criterion of factors with eigenvalues greater than 1.0 was used for screening (Field, 2005; Meyers, Gamst, & Guarino, 2006; Stevens, 2002). To enhance interpretation of components meeting Kaiser's criterion, scree plots were analyzed.

6. With regard to the principal component analysis, critical values of factor loadings for various sample sizes were provided by Stevens (2002), beginning with 50 samples. For n-sizes smaller than 50, Stevens recommended using the standard error doubled as a criterion for rejection.

7. As a follow-up to principal component analysis, Cronbach's α test was conducted. Cronback's α reliability within the component should be greater than 0.7 (Spector, 1992). As Cronbach's α is capable of processing dichotomous as well as interval scores, unlike Kuder-Richardson's tests, it was the preferred measure of reliability. It is noted that in the case of smaller sample sizes and fewer variables, Cronbach's α is sensitive to numbers of variables. Smaller numbers of items usually result in smaller α's.

Summary

This chapter briefly restated the research questions and characterized the research design as exploratory, *ex post facto*, using a mixed model of qualitative and quantitative data received from interviews and surveys of participant teachers as well as statistical applications. The data was to be coded and analyzed using principal component analysis. Criteria for use with the statistical processes were specified. Reliability testing was to be accomplished on the components revealed using Cronbach's ∀ and the contributing variables examined with the goal of functionally naming the components revealed.

In Chapter IV, the findings of the study will be reported. Specifically, the processes of final subject selection and study responses are detailed. Instrumentation usage and data coding are briefly discussed. The preliminary analyses to which the collected data were subjected are reported as well as the results. Each of two Research Questions is then examined and findings of the principal component (factor) analysis reported.

CHAPTER IV

FINDINGS

The goal of this study was to examine the contribution that education and training in selected topics made to a teacher's capacity to use formative testing results to inform instruction in third grade mathematics and thereby increase learner achievement. The research was intended to reveal which specific variables drawn from education and training histories, if any, contributed to factors or components that might lead to the improvement of her/his capacity to use formative assessment to guide her/his teaching. Variables included beliefs, content and focus of teacher preparation programs, experience in the classroom, the focus and content of graduate programs, and selected professional development and training. All variables could contribute to the constructs underlying a teacher's effective use of formative testing to inform instruction but to varying degrees. To these ends, research was conducted following the granting of human subjects' exemption and a favorable review of the proposed research by the school district hosting the study. Data were collected and analyzed. This chapter presents the findings of that research.

This chapter begins with an accounting of response to the study including the data collected. The stratification scheme for the data collected is examined as a means of making relative sense of the data and the components subsequently revealed. Results of the data collection instrument's reliability analyses are provided. The coding of data collected is briefly discussed while a preliminary analysis of what these data suggest, when stratified by participant school performance, is offered. The results of the principal component analysis conducted are presented. Finally, the research questions are

examined in light of the resulting components. The chapter is then summarized.

Study Responses

Three quarterly mathematics tests were administered to an average of 2575 third grade students attending classes in 163 classrooms of 35 elementary schools within the district during school year 2004-2005. Schools were ranked based on their composite scores using the Colorado Student Assessment Program (CSAP) model (Linn & Haug, 2002). The CSAP model was devised to reduce the instability often seen when using scores as indices of achievement. Instead, ordinal groups of advanced, proficient, partially proficient, and below partially proficient are used in a weighted formula to rank schools for accreditation. The district's elementary schools were ranked accordingly. The design created three school groups comprised of classrooms (i.e., teachers) representing high achievement, median achievement, and low achievement by the included third grade students. These groups were designated Groups A, B, and C, respectively.

To preclude confounding influences of schools using different instructional and testing models (i.e., different foci of teacher curricular and testing preparation and application), seven schools were noted to employ other than classic instructional models (e.g., departmentalized instruction), five schools used school-wide testing more often than the district assessment plan, and one school employed both. These 13 schools were removed from consideration by this study. Principals of the 22 remaining schools were invited to participate in the study. Three opted not to participate, one school was discovered to have been misclassified by both instructional and testing models during the interview process, and one school did not respond. The remaining 17 schools and 78 assigned teachers constituted the final, accessible population, a population of

convenience, from which a sample might be drawn and data might be collected.

Preliminary examination of teacher participation, whether by interview or survey, revealed four groupings with low or non-existent participation rates. The largest group was comprised of self-contained special education teachers. On further examination, most of these teachers' students were fourth and fifth grade students who had been given third grade mathematics assessments in accordance with those students' individual education plans (IEPs). Other low or non-existent response groupings included teachers who had retired, teachers who had relocated outside the Commonwealth of Virginia, and teachers reassigned within the school district. These groupings and teacher counts are listed in Table 7.

Table 7
Teachers Not Considered in Study, by Groupings

Teacher Groupings	Number
Self-contained special education (IEP associated use of third grade tests)	10
Retired	5
Relocated outside of Virginia and opted not to participate	4
Reassigned within district and opted not to participate	3

Of the 56 teachers remaining, 41 provided data during interview or by survey. Three special education teachers and two teachers reassigned within the district contacted directly by mail (pre-paid postage survey) late in the data collection phase, opted to participate resulting in 46 total respondents. Thirty-two (69.6%) subjects provided data by interview while 14 (30.4%) participated via mailed survey. The 46 participants represented 75.4% (46 of 61, 56 possible plus 5 unexpected participants) of accessible teachers.

During preliminary stages of data collections, 15 schools were contacted. The

early weeks of data collection efforts were slow. In order to obtain more data, the two extra and all previously non-participant eligible schools were contacted/re-contacted at the six week point. Over the ensuing weeks, data were collected from all 17 schools. Four schools had 100% participation rates. At least one of these schools was in each of the three school groupings.

With 17 schools responding, the original stratification plan using five schools in each of three groupings (5 x 3) was exceeded though skewed in favor of higher performing schools. Data were collected from 13 teachers and six schools of Group A, 22 teachers and seven schools in Group B, two teachers and one school constituting a strata boundary school, and 9 teachers and three schools in Group C. In an effort to balance the number of teachers and schools in each group, the strata boundary school between original groups B and C and the lowest performing school and its teachers in Group B were reassigned to Group C bringing Group C to 15 teachers and five schools. Statistically, the reassignment appeared justified as these schools were similar to other schools in Group C. To maintain the planned schema of 5 x 3, two schools, the lowest achieving schools in Group A and revised Group B were omitted resulting in five schools in each of three strata. Data for three responding special education teachers were also omitted from school groupings resulting in final groups of 11, 14, and 14 teachers, respectively, in Groups A, B, and C. The data supporting the distribution of the 17 responding schools including Colorado Student Achievement Program ranking, Colorado Student Achievement Program scores, original and final groupings, and participating teacher counts are displayed in Table 8. The teacher *n*-size for examination of group strata was 39. For the purposes of principal component analysis, however, all available

teacher data were used. The *n*-size for principal component analysis was 46.

Table 8
Participating Schools' Distribution

CSAP Rank	CSAP Score	Original Group	Final Group	Participating Teacher Count
1	0.6824	A	A	2
2	0.6759	A	A	2
6	0.5639	A	A	2
9	0.5094	A	A	2[a]
10	0.4942	A	A	3[a]
11	0.4687	A	SB[b]	2
13	0.4388	B	B	4[c]
14	0.3833	B	B	4
15	0.3789	B	B	1
17	0.3711	B	B	3
19	0.3624	B	B	2[c]
22	0.3242	B	SB[b]	2
23	0.2931	B	C[d]	4
24	0.2920	SB	C[d]	1[c]
26	0.2452	C	C	4
27	0.2295	C	C	3
34	0.1354	C	C	2

Note. [a]includes one within district reassignment. [b]strata boundary, not included in comparison of stratified data. [c]does not include one unexpected response from a special education teacher. [d]reassigned to Group C.

Instrumentation and Data Coding

The protocol for this study addressed information relevant to the goals of the research questions. As the study was exploratory, there were several aspects that emerged during the study that had not been *a priori* considerations. The ways in which emergent data were considered are herein described.

For the instrument used, items with dichotomous and interval or ordinal values were tested, after standardization, with Cronbach's α. These data included 11 items associated with training and 17 items associated with education. Using a thumb rule

provided by Nunnally (1978, as cited by Spector, 1992) suggesting that at the level of 0.7 internal consistency is acceptably demonstrated, all but two item pairs of dichotomous/interval data had Cronbach's α's above 0.7. Item pairs addressing the topic of assessment in undergraduate classes achieved a 0.674, while one addressing differentiated instruction training achieved a 0.559. Complete results, including response counts, missing data, means, standard deviations, Pearson's r, and Cronbach's α with F-Test and significance, are presented in Table 9.

Data collected appeared to have a time-related bias. This aspect was considered a potential limitation and is based on the clarity of teachers' recollections. For some teachers, recall of hours spent studying a particular topic occasionally seemed ambiguous. That she/he had been introduced to the topic at a discrete stage of her/his respective educations was recorded with more certainty. More recent education or training events resulted in more precise data. Review of records did fill-in some areas of weak recollection, but not all.

While not originally part of the study, it was evident in interviews that teachers generally trusted specific individuals within their respective schools for information concerning analysis of testing data, revision of instruction, and formulation of strategies. These opinion leaders (Rogers, 1995), those persons to whom one goes with a problem, consistently appeared to be current or prior special education teachers. It is recognized that special education teachers are education graduates; they have been, however, schooled with the recognition of the requirement to treat the specific needs of each student as an individual. While transcribing degree major data, a pattern emerged. A rank ordered field was created based on teacher responses. One end of the continuum was

Table 9

Teacher Interview and Survey Protocol Item Statistics

Pair	Item Code	Question Stem	Responses	Missing Data	Mean	S.D.	Pearson's *r*	Cronbach's α (standardized)	F-Test	Sig.
1	B01	Did you have mathematics courses beyond the requirements of your degree program?	Yes – 24 No – 22	5	0.522	0.505	0.778	0.808	24.920	0.000
	Q01	If so, what were they?	21	3	50.43	58.81				
2	B02	Did you have any course(s) in which assessment was a topic of discussion?	Yes – 31 No – 15		0.674	0.474	0.573	0.674	35.080	0.000
	Q02	If so, do you recall how much class-time was spent on it? (hours)	23	8	17.70	17.22				
3	B03	Did you have any course(s) in which testing was a topic of discussion?	Yes – 32 No – 14		0.696	0.465	0.555	0.713	28.941	0.000
	Q03	If so, do you recall how much class-time was spent on it? (hours)	24	8	16.42	16.02				
4	B04	Did you have any course(s) in which item writing was a topic of discussion? (hours)	Yes – 16 No – 30		0.348	0.482	0.599	0.749	7.937	0.007
	Q04	If so, do you recall how much class-time was spent on it? (hours)	10	6	17.10	24.61				
5	B05	Did you have any course(s) in which data-driven instruction was a topic of discussion? (hours)	Yes – 13 No – 33		0.283	0.455	0.561	0.722	4.651	0.036
	Q05	If so, do you recall how much class-time was spent on it? (hours)	10	3	7.60	11.42				
6	B07	Have you had graduate classes that contained evaluation?	Yes – 22 No – 23		0.500	0.506	0.606	0.932	33.763	0.000
	Q06	Hours?	18	5	40.56	43.83				
7	B08	Have you had graduate classes that contained statistics?	Yes – 20 No – 26		0.435	0.501	0.956	0.977	31.469	0.000
	Q07	Hours?	17	3	36.35	9.38				
8	B09	Have you had graduate classes that contained testing?	Yes – 21 No – 25		0.457	0.504	0.830	0.907	25.421	0.000
	Q08	Hours	19	2	28.16	15.08				
9	B10	Have you had graduate classes that contained test item construction?	Yes – 14 No – 32		0.304	0.465	0.816	0.898	12.754	0.001
	Q09	Hours?	12	2	26.25	17.49				

Table 9 - continued

Teacher Interview and Survey Protocol Item Statistics

Pair	Item Code	Question Stem	Responses	Missing Data	Mean	S.D.	Pearson's *r*	Cronbach's α (standardized)	F-Test	Sig.
10	B11	Have you had graduate classes that contained differentiated instruction?	Yes – 26 No – 20		0.565	0.501	0.644	0.784	23.591	0.000
	Q10	Hours?	22	4	38.64	33.64				
11	B12	Have you had graduate classes that contained research design?	Yes – 26 No – 20		0.565	0.501	0.822	0.903	38.862	0.000
	Q11	Hours?	25		39.00	16.01				
12	B13	Have you had graduate classes that contained assessment?	Yes – 25 No – 21		0.543	0.504	0.825	0.904	35.743	0.000
	Q12	Hours?	23	3	33.57	16.58				
13	B14	Have you had graduate classes that contained standards?	Yes – 12 No – 34		0.261	0.444	0.786	0.880	13.168	0.001
	Q13	Hours?	11	2	28.73	15.75				
14	B15	Have you had graduate classes that contained analysis of results?	Yes – 16 No – 30		0.348	0.482	0.817	0.899	15.672	0.000
	Q14	Hours?	14	2	31.71	20.08				
15	B16	Have you had graduate classes that contained advanced curriculum design?	Yes – 18 No – 28		0.391	0.493	0.797	0.887	18.025	0.000
	Q15	Hours?	17	1	41.12	25.82				
16	B17	Have you had graduate courses that contained additional mathematics content?	Yes – 20 No – 26		0.435	0.501	0.787	0.881	21.149	0.000
	Q16	Hours?	19	1	58.67	32.81				
17	B18	Do you have graduate courses beyond your master's degree?	Yes – 11 No – 35		0.239	0.431	0.845	0.916	10.039	0.003
	Q17	Hours?	11		99.55	57.64				
18	B19	Have you had additional training in math methods?	Yes – 44 No – 2		0.957	0.206	0.598	0.749	208.729	0.000
	O01	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours		3.250	1.120				

Table 9 - continued

Teacher Interview and Survey Protocol Item Statistics

Pair	Item Code	Question Stem	Responses	Missing Data	Mean	S.D.	Pearson's *r*	Cronbach's α (standardized)	F-Test	Sig.
19	B20	Have you had additional training in assessment?	Yes – 42 No – 3		0.935	0.250	0.571	0.728	116.202	0.000
	O02	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours	3	3.128	1.379				
20	B21	Have you had additional training in testing?	Yes – 42 No – 4		0.913	0.285	0.645	0.784	114.223	0.000
	O03	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours	2	3.025	1.366				
21	B22	Have you had additional training in test results analysis?	Yes – 42 No – 4		0.913	0.285	0.652	0.789	118.294	0.000
	O04	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours	1	3.049	1.347				
22	B23	Have you had additional training in data-driven decision-making?	Yes – 43 No – 3		0.935	0.250	0.600	0.750	132.850	0.000
	O05	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours		3.128	1.297				
23	B24	Have you had additional training in action research?	Yes – 15 No – 31		0.326	0.474	0.902	0.949	13.585	0.001
	O06	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: 0 hours		2.600	1.370				
24	B25	Have you had additional training in evaluation?	Yes – 35 No – 11		0.761	0.431	0.803	0.890	61.346	0.000
	O07	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours	2	2.818	1.540				
25	B26	Have you had additional training in differentiated instruction?	Yes – 45 No – 1		0.978	0.147	0.388	0.559	146.746	0.000
	O08	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours		2.978	1.132				

Table 9 - continued

Teacher Interview and Survey Protocol Item Statistics

Pair	Item Code	Question Stem	Responses	Missing Data	Mean	S.D.	Pearson's *r*	Cronbach's α (standardized)	F-Test	Sig.
26	B27	Have you had additional training in strategies, activities, and exercises?	Yes – 44 No – 2		0.957	0.206	0.592	0.743	208.000	0.000
	O09	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours		3.364	1.172				
27	B28	Have you had additional training in instructional planning and revision?	Yes – 38 No – 8		0.826	0.383	0.800	0.889	98.255	0.000
	O10	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: > 10 hours		3.237	1.550				
28	B29	Do you conduct professional development, in-service, or capacity development training as an instructor?	Yes – 15 No – 31		0.326	0.474	0.937	0.968	16.912	0.000
	O11	0 hours, 2 or less hours, 2 to 5 hours, 5 to 10 hours, more than 10 hours	Mode: 0 hours							

defined by education while the other end represented degrees in subjects not normally associated with elementary education (e.g., business administration, British literature, information systems, etc.). Of those with education degrees and, based on the anecdotal observations of the teachers interviewed, special education teachers (e.g., speech pathology, deaf education, etc.) were established as a separate category of education graduates. Minors noted were used to shade these rankings. The five possible ordinal field entries were special education or populations, educational programs, psychology and sociology, content specialization, and other. These entries and their ranks are listed in Table 10.

Table 10
Degree Program Majors Coding

Major	Rank
Special education/special populations (e.g., deaf education)	1
Educational programs (e.g., early childhood)	2
Psychology/sociology	3
Content specialists (e.g., biology, British literature)	4
Other (e.g., business administration, information systems)	5

Preliminary Analyses

Preliminary analyses of the data collected were conducted for a number of reasons. First, Field recommends data screening in order to determine whether the variables to be analyzed are sensible thus preventing a "if you put garbage in, you get garbage out" scenario (2005, p.640). Second, the examination of emergent patterns potentially lends explanation or confirmation to components extracted. Finally, examination of patterns might suggest ways in which the three stratified school groupings were different, potentially lending greater understanding of components revealed. Data suggesting why some schools performed better than others were examined. The examination of these data

as possible sources of difference and potential contributions to the components revealed also provided the opportunity to consider their respective merits in the process at large. In the examination of the data and their relevance to achievement for this study, there are four discrete groupings for consideration: environment, undergraduate education, graduate education, and professional development and training. Using school groupings identified as Group A – high performing schools, Group B – median performing schools, and Group C – low performing schools, the average achievement by school groupings, schools, and classrooms (i.e., teachers) that participated in this study were examined. Group A schools averaged (72.5%) more than four percentage points higher than Group B schools (68.3%), while Group B schools averaged more than four percentage points higher than Group C schools (63.6%). These data and school averages are presented in Table 11. These data are listed in CSAP rank order though participant school averages may not appear to reflect that order.

Table 11
School and Classroom Achievement

Group A Average:		72.5%	Group B Average:		68.3%	Group C Average:		63.6%
School	Classrooms	Average	School	Classrooms	Average	School	Classrooms	Average
1	2	78.5%	1	4	68.7%	1	4	65.6%
2	2	73.4%	2	4	66.4%	2	2	59.9%
3	2	72.7%	3	1	74.7%	3	4	61.2%
4	2	68.6%	4	3	68.2%	4	3	65.3%
5	3	70.2%	5	2	71.1%	5	2	63.6%

Note. The average scores presented are averages for the students of teachers that participated in the study and not the ranking achieved by the school using the CSAP model score, hence the apparent disconnect in ranking of scores listed.

One note, throughout this section, several comparisons neared but did not achieve significance. Whether failure to be significant was due to small sample sizes or truly non-significant differences, the *p*-values are offered as those comparisons may be worthy of

re-examination at a later time with larger sample sizes.

Class Size and Teacher Experience

Environment normally includes any consideration that potentially contributes or detracts from a process based on surroundings. According to the Oregon Public Education Network (2004), environment establishes a place, time, and atmosphere reflective of conditions that affect thoughts and actions. The effects of class size are often a consideration in achievement (Darling-Hammond & Ball, 1998). The average class size for all groups seems nearly equivalent with the largest associated with Group A schools at 20.7 students compared to an all-schools average of 19.3. However, the mode of Group A classrooms (6) was 22-23 students with an average of 20.73 while the modes for Group B (5) and C (5) schools were the same, 18-19 students with averages of 18.29 and 19.29, respectively. The difference in Group A and Group B schools' class size was significant ($p = 0.035$). The difference between Groups A and C was not significant ($p = 0.120$). Specific data, by group, may be found in Table 12. The range of class sizes, by group and size, are shown in Figure 4.

Table 12
Class Size and Teacher Experience

	All	Group A		Group B		Group C	
	Ave.	Rank	Ave.	Rank	Ave.	Rank	Ave.
Class size (students)	19.3	3	20.7[a,b]	1	18.3[a]	2	19.3[b]
Years teaching third grade	6.2	2	6.6	3	5.1	1	6.9
Total years teaching	12.6	1	17.6[c,d]	3	10.5[c]	2	10.9[d]

Note. [a]Differences between Groups A and B were significant ($p = 0.035$). [b]Differences between Groups A and C were not significant ($p = 0.120$). [c]Difference between Groups A and B were not significant ($p = 0.072$). [d]Differences between Groups A and C were not significant ($p = 0.135$).

Overall, the 39 teachers in the three stratified groups averaged 12.6 years of total

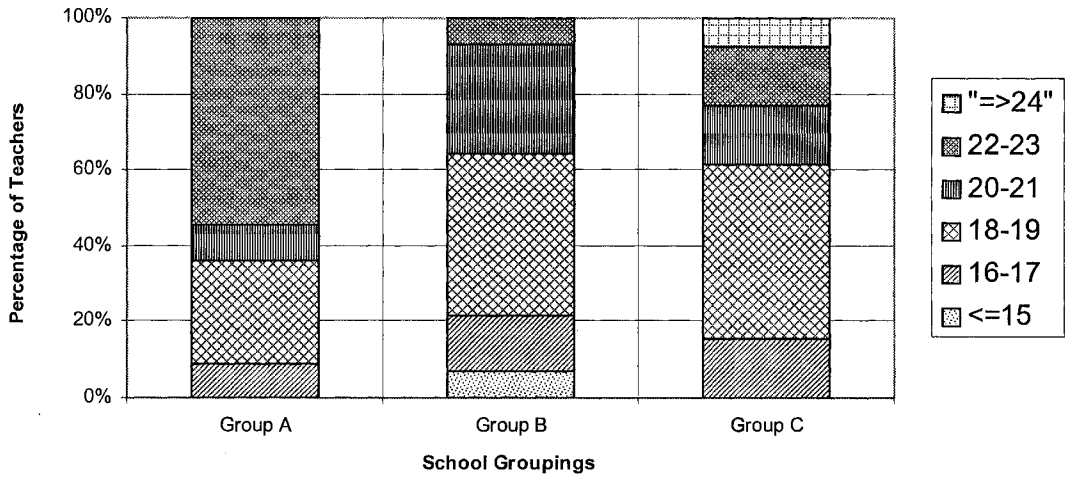

Figure 4. Class size comparison in two year intervals.

teaching experience (SD =10.0) with an average 6.2 years teaching third grade (SD = 4.9). However, Group A teachers averaged 17.6 years total teaching experience (SD = 10.9) while Groups B and C averaged 10.5 and 10.9 years (SD = 7.8, SD = 10.6), respectively. Though seemingly substantial, the difference between Groups A and B was not significant ($p = 0.072$). Regarding experience teaching third grade, Groups A (6.6 years) and C (6.9 years) were nearly equivalent while Group B teachers averaged 5.1 years. Group B teachers had the lowest average number of years teaching experience in both categories. These data are also presented in Table 12. Graphical representations of years of experience, total and in third grade, are presented in Figures 5 and 6, in five year intervals which show the variations and ranges of experience encountered.

Undergraduate Educational Foci

Schools in Groups A and B had a slight advantage in percentage of teachers with education as an undergraduate focus with 54.6% and 57.1%, respectively, of all teachers studying in programs pertinent to either education or special education. Group C schools had 50.0%. In fields of psychology and sociology, Groups A and B had 18.2% and 21.4%,

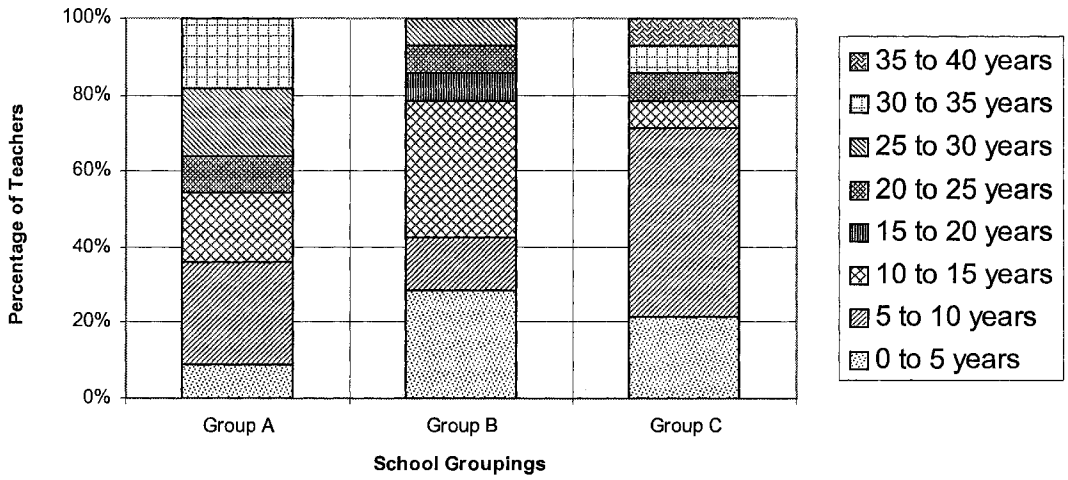

Figure 5. Overall years of teaching experience, in five year increments, within school groupings A, B, and C.

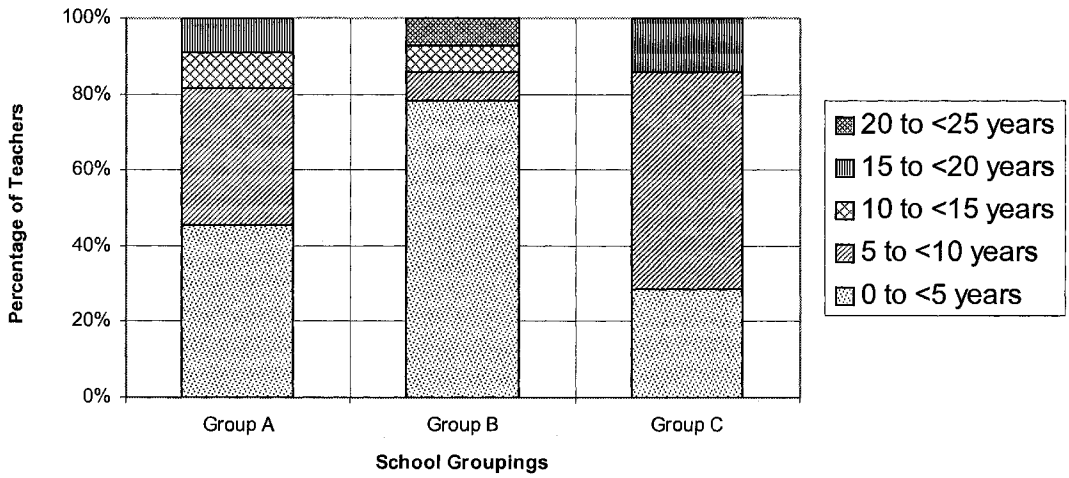

Figure 6. Years experience teaching third grade, in five year increments, within school groupings A, B, and C.

respectively, of graduates while Group C had 7.1%. In areas that might be considered a content area undergraduate foci (i.e., French, British literature, biology, mathematics, art, art history, and American studies), the opposite trend was noted. Group A and B teachers had 9.1% and 7.1%, respectively, while Group C had 35.7% which represented five of seven teachers in this category. For program foci in other areas (i.e., business administration, management information systems, and economics), Groups A, B, and C

had 18.2%, 14.3%, and 7.1%, respectively, of their graduates in these fields with one business administration graduate in each group. All totaled, 6 of 12 teachers with undergraduate degrees in areas other than education or education related fields (art, French, British literature, biology, mathematics, and business administration) were assigned to Group C schools where they comprised 42.86% of Group C teachers. The remaining six teachers with non-education related degrees were evenly distributed to Group A (business administration, economics, and art history) and Group B (American studies, business administration, and management information systems) schools where they accounted for 27.3% and 21.4% of assigned teachers, respectively. These data are presented as percentages in Table 13, and as counts and rankings in Table 14. The complete distribution of teachers' undergraduate degrees, by area of study, is depicted in Figure 7.

Table 13
Teachers' Undergraduate Educational Foci by percentages

	All	Group A	Group B	Group C
Special education fields	7.7%	18.2%	7.1%	0.0%
Education	46.2%	36.4%	50.0%	50.0%
Psychology/sociology	15.4%	18.2%	21.4%	7.1%
Content	17.9%	9.1%	7.1%	35.7%
Other	12.8%	18.2%	14.3%	7.1%

Graduate Educational Foci

In the examination of graduate or continuing education, 89.7% of subjects had taken graduate level courses. Twenty of the 39 teachers (52.3%) had earned master's degrees prior to the conclusion of this study. Teachers with master's degrees were near evenly distributed with six in Group B schools and seven each in Group A and C schools where they comprised 42.9%, 63.6%, and 50.0% of all teachers, respectively. All

Table 14
Teachers' Undergraduate Educational Foci by numbers and rank

	Group A		Group B		Group C	
	Rank	Teachers	Rank	Teachers	Rank	Teachers
Special education	1	2	2	1	3	0
Education	3	4	1	7	1	7
Psychology/sociology	2	2	1	3	3	1
Content	2	1	2	1	1	5
Other	1	2	1	2	3	1

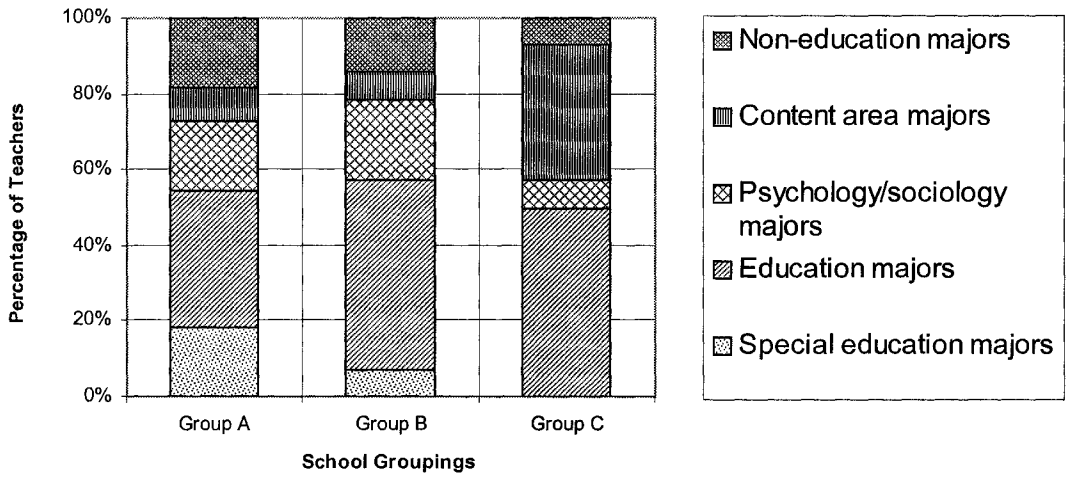

Figure 7. Teachers' undergraduate degrees grouped by emergent fields of study within school groupings A, B, and C.

master's degrees were in education (16), special education (2), or school administration (2). While not directly queried during data collection, based on the proximity of master's degree award dates to the commencement of teaching and the comparison of bachelor's and master's degree foci, 9 of 20, or 45.0%, of master's degrees appeared to have been awarded in preparation for licensure. Four of these likely preparatory master's degrees were held by Group C teachers, three by Group A teachers, and two by Group B teachers. Of master's degrees earned after initial licensure, four each were held by Group A and B teachers while three were held by Group C. The distribution of graduate degrees is shown

in Figure 8. The distribution of differences between Group A and Group C schools, based on area of study, was not statistically significant ($p = 0.184$). These data are summarized in Table 15. It was noted that the average degree award years for Groups A, B, and C were 1992, 1998, and 1997, respectively. This distribution appears to provide Group A teachers with a five year plus advantage with the content knowledge acquired over Groups B and C.

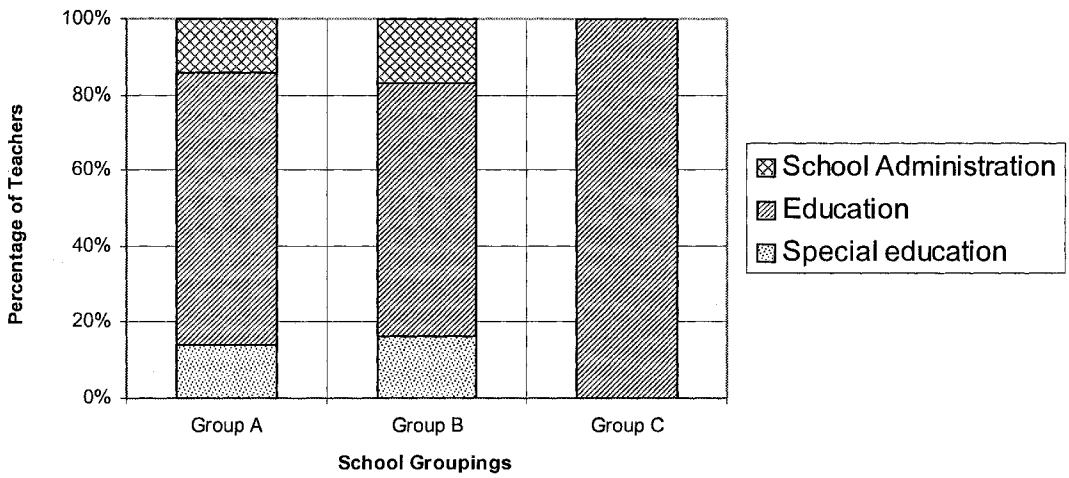

Figure 8. Teachers' graduate degrees grouped by area of study within school Groups A, B, and C.

Table 15
Teachers' Graduate Educational Foci and Stage

	Group A		Group B		Group C	
	Rank	Teachers	Rank	Teachers	Rank	Teachers
Special education	1	1	1	1	3	0
Education	2	5	3	4	1	7
School administration	3	1	1	1	1	0
Preparatory programs[a]	2	3	3	2	1	4
Post-preparatory programs[b]	1	4	1	4	3	3

Note. [a]Preparatory programs are defined here as those master's degree programs that complete requirements for licensure. [b]Post-preparatory programs are defined here as those master's degree programs taken post-licensure and pertain to a classic mastery of program content.

Regarding comparisons of undergraduate foci and master's programs, 12 of 20

master's degrees earned were awarded to those with undergraduate foci in education or special education programs for 57.1%. In these foci, 66.7%, 50.0%, and 28.6%, respectively, of Group A, B, and C teachers earned master's degrees. In the groupings of psychology and sociology, content areas, and non-educational foci, 60.0%, 20.0%, and 71.4% of Group A, B, and C teachers earned master's degrees in education. Of note, seven of nine master's degrees earned in preparation for licensure, were awarded to teachers with non-educational foci bachelor's degree programs for a total of 41.2% of non-educational undergraduates. Three of Group A's five, one of Group B's five, and three of Group C's seven non-educational undergraduates earned master's degrees in education in association with licensure. Of the 19 teachers without master's degrees, 36.4%, 57.1%, and 50.0% were assigned to Groups A, B, and C, respectively. For teachers with non-educational undergraduate foci and no master's degree, 40% were part of Group A, 80% assigned to Group B, and 28.6% assigned to Group C schools.

Likes, Beliefs, and Assistance

Data regarding teachers' personal likes or preference for mathematics, belief in formative testing as practiced by the district, and the estimation of the support received from respective mathematics specialists were recorded as Likert-scaled items. While these topics are represented by single items, they potentially provide insights into topics discussed in the literature.

Teachers in Groups A and C were largely positive in their personal *like* or preference for mathematics with 81.8% and 85.6%, respectively. Group A teachers' preference or *like* was at worst neutral while the remainder of Group C teachers (2) held neutral or negative sentiments. While the majority of Group B schools' teachers were

positive, 5 of 14 (35.7%) teachers held neutral or negative sentiments. The differences in Group A teachers' *like* (i.e., preference) for mathematics compared to Group B teachers was not significant ($p = 0.135$). Percentages of these data are recorded in Table 16 and charted in Figure 9.

Table 16
Teachers' Preferences, Beliefs, and Evaluations and Mathematics Curriculum Coordinator's Evaluations

				Most	5-point Likert-scale			Least
		Rank	Composite Score[a]	1	2	3	4	5
Like of (preference for) mathematics	Group A	1	1.73[b]	45.5%	36.4%	18.2%	0.0%	0.0%
	Group B	3	2.36[b]	21.4%	42.9%	21.4%	7.1%	7.1%
	Group C	2	1.79	57.1%	28.6%	0.0%	7.1%	7.1%
Belief in formative program	Group A	1	1.91[c]	27.3%	54.5%	18.2%	0.0%	0.0%
	Group B	3	2.64[c]	0.0%	57.1%	28.6%	7.1%	7.1%
	Group C	2	2.36	28.6%	21.4%	35.7%	14.3%	0.0%
Teachers' mathematics specialist's evaluation	Group A	2	2.91[d]	36.4%	18.2%	0.0%	9.1%	36.4%
	Group B	3	3.64[e]	7.1%	14.3%	21.4%	21.4%	35.7%
	Group C	1	1.71[d,e]	64.3%	14.3%	14.3%	0.0%	7.1%
Mathematics coordinator's mathematics specialist evaluation	Group A	3	2.45[f]	18.2%	45.5%	18.2%	0.0%	18.2%
	Group B	2	2.21	42.9%	21.4%	7.1%	28.6%	0.0%
	Group C	1	1.57[f]	78.6%	0.0%	7.1%	14.3%	0.0%

Note. [a]Composite score is the weighted average of responses within a group. [b]Differences between Groups A and B are not significant ($p = 0.135$). [c]Differences between Groups A and B are significant ($p = 0.040$). [d]Differences between Groups A and C are not significant ($p = 0.084$). [e]Differences between Groups B and C are significant ($p < 0.000$). [f]Differences between Groups A and C are not significant ($p = 0.066$).

Analysis of teachers' beliefs in the district's formative assessment as practiced found that all Group A teachers were either neutral (2 – 18.2%) or positive (9 – 81.8%) while schools in Groups B and C had more teachers who were either neutral or negatively disposed. By comparison, Group B had no teachers who strongly believed in the program while 2 of 14 held negative or strongly negative beliefs. This difference in beliefs between Groups A and B was statistically significant ($p = 0.040$). In Group C, 50% of

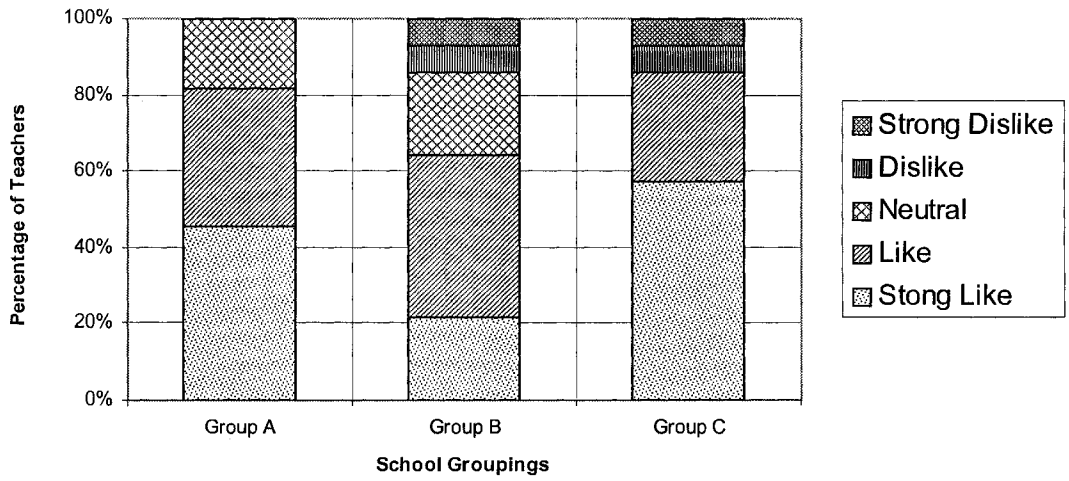

Figure 9. Teachers' personal "like" for mathematics reported on a 5-point Likert-scale within school groupings A, B, and C.

teachers were positive in their beliefs while 50% were neutral or negative. These data are also presented in Table 16 and are charted in Figure 10.

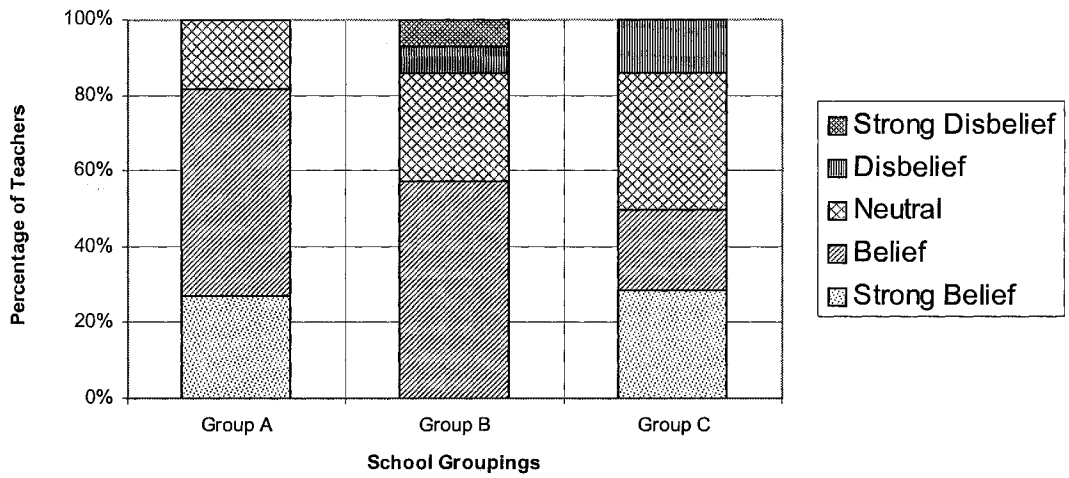

Figure 10. Teachers' belief in formative assessment as practiced in the district within school groupings A, B, and C, as reported using a 5-point Likert-scale.

The examination of teachers' and mathematics curriculum coordinator's evaluations revealed a variety of results. Regarding teachers' inputs, Group A had a bimodal distribution with 36.4% of teachers reporting the greatest and 36.4% reporting the least with no teacher being neutral. Group B teachers' opinions were increasingly

skewed or ramped toward negative evaluations with only 6 of 14 being positively or neutrally disposed. By comparison, Group C teachers were largely satisfied with the assistance received from their mathematics specialists with 78.6% reporting high or moderately positive evaluations. The difference between Group B and Group C teachers was significant ($p < 0.000$) while the difference between Group A and C teachers was not ($p = 0.065$). The mathematics curriculum coordinator's evaluations of mathematics specialists differed from the perceptions of teachers. Group B and C schools' mathematics specialists were generally evaluated more positively by the mathematics curriculum coordinator than by teachers. Examination of correlations between teachers' and mathematics curriculum coordinator's evaluations revealed an overall correlation of $r = 0.254$, and within group correlations of 0.608, -0.656, and 0.567 for Groups A, B, and C, respectively. These data, in percentages, are presented in Table 16 and Figure 11.

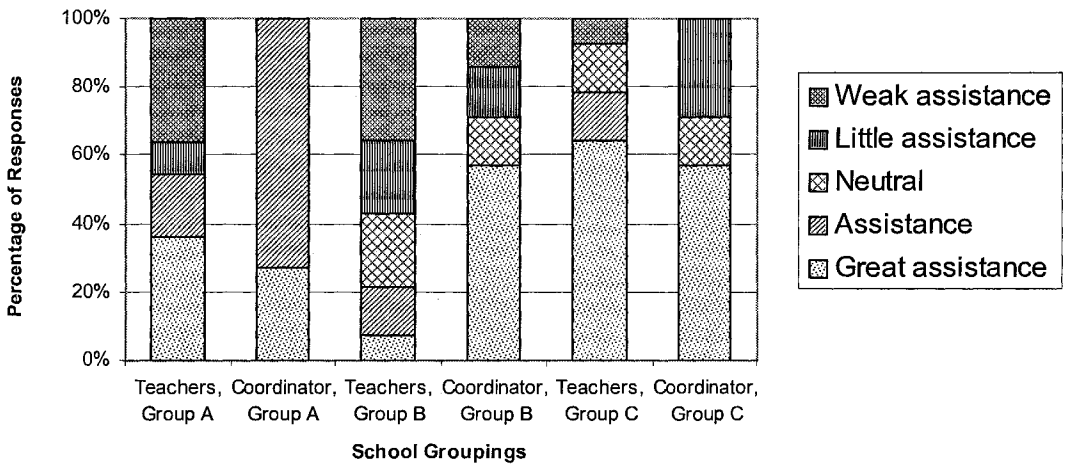

Figure 11. Evaluation of schools' mathematics specialist assistance as reported on a 5-point Likert-scale by teachers and mathematics curriculum coordinator within school groupings A, B, and C.

Undergraduate Topic Exposure

The examination of undergraduate topic exposure revealed a number of

differences among school groupings. Regarding mathematics courses beyond licensure requirements, 64.3% of Group C teachers reported additional instruction while only 42.9% of Group B and 36.4% of Group A teachers did so. Review of the number of hours of additional instruction reported, however, revealed that Groups A, B, and C reported averages of 14.0, 11.2, and 9.6 hours, respectively. The differences between Groups A and C were not significant ($p = 0.104$).

In areas of assessment and testing as topics in other courses, 81.8% of Group A teachers recalled instruction in both of these topics for averages of 16.5 and 8.0 instructional hours, respectively. Group B teachers' percentages were slightly less with 78.6% and 71.4% recalling participation (9.8 and 6.2 hours), while 42.9% and 50.0% (11.9 and 4.0 hours) of Group C teachers recalled instruction in these topics. The difference in the recollections of assessment as a topic, however, was found to be statistically significant between Groups A and C ($p = 0.045$) but not significant between Groups B and C ($p = 0.056$). Differences between Groups A and C, regarding the recollection of testing instruction, were not significant ($p = 0.098$).

Fewer than one-third of all teachers recalled instruction in test-item writing at the undergraduate level. For those that did, an average of 3.7 hours of instruction was noted.

Data-driven instruction as a topic was reported by about one-fourth of teachers but for those that did, no fewer than 25 hours of instruction were reported. Half of Group C teachers accumulated an average number of hours in data-driven instruction equivalent to more than a course length exposure. There were no significant differences between Groups. The recollections of undergraduate education topics in courses and the number of hours reported, as modified by record content review, are tabulated in Table 17.

Table 17
Teachers' Recollection of Select Undergraduate Educational Topics, Hours of Instruction with Correlation and Reliability Data

	Pearson's *r* and Cronbach's ∀ of recollection to instructional hours		Recalled Instruction (percentage)			Instructional Hours (averages for those taken)		
	r	∀	Group A	Group B	Group C	Group A	Group B	Group C
Total teachers			11	14	14			
Mathematics beyond program	0.778	0.808[g]	36.4%	42.9%	64.4%	14.0[e]	11.2	9.6[e]
Assessment	0.573	0.674[h]	81.8%[a]	78.6%[b]	42.9%[a,b]	16.5	9.8	11.9
Testing	0.555	0.713[i]	81.8%[c]	71.4%	50.0%[c]	8.0	6.2	4.0
Item writing	0.599	0.749[j]	18.2%	42.9%	21.4%	3.0[f]	3.4[f]	4.3
Data-driven instruction	0.561	0.722[k]	9.1%[d]	35.7%	28.6%[d]	27.4	34.2	43.0

Note. [a]Differences between Groups A and C are significant ($p = 0.045$). [b]Differences between Groups B and C are not significant ($p = 0.056$). [c]Differences between Groups A and C are not significant ($p = 0.098$). [d]Differences between Groups A and C are not significant ($p = 0.113$). [e]Differences between Groups A and C are not significant ($p = 0.104$). [f]Differences between Groups A and B are not significant ($p = 0.093$). [g]Significant ($p < 0.000$). [h]Significant ($p < 0.000$). [i]Significant ($p < 0.000$). [j]Significant ($p = 0.007$). [k]Significant ($p = 0.036$).

Graduate Course Exposure

The topics of interest at the graduate level were typically those that might have a dedicated course or were part of courses dedicated to a few related topics (e.g., testing, test-item writing, and testing results analyses). Based on instructional hours accumulated, it was noted that differentiated instruction and test-item construction were the leading topics with all groups reporting an average of more than 30 hours of instruction for each teacher who participated. It was noted that the recollections of taking differentiated instruction as a course favored school groupings in rank order, i.e., Groups A - 63.6%, B - 57.1%, and C - 42.9%, although the hours accumulated averaged about the same (34.0, 33.7, and 36.6, respectively). Courses in evaluation, statistics, assessment,

standards, and analyses appeared to favor Group A, based on recollection of taking a course, but the hours reported favored Groups B and C. Only Groups B and C had significant differences in the number of hours reported in advanced curriculum design ($p = 0.038$). Percentages for recalled instruction and hours reported for selected graduate topics are reported in Table 18.

Table 18
Teachers' Recollections and Average Hours of Graduate Instruction in Select Topics with Correlation and Reliability Data

	Pearson's *r* and Cronbach's ∀ of recollection to instructional hours		Recalled Instruction (percentage)			Instructional Hours (averages for those taken)		
Topics taken as graduate courses	r	∀	Group A	Group B	Group C	Group A	Group B	Group C
Total teachers			11	14	14			
Evaluation[g]	0.606	0.932[a]	72.7%[c]	42.9%[c]	50.0%	31.0	38.8	37.8
Statistics	0.956	0.977[a]	54.5%	42.9%	42.9%	17.0[d]	31.5	28.5[d]
Testing[f]	0.830	0.907[a]	36.4%	35.7%	57.1%	9.5	24.8	29.5
Test item writing	0.816	0.898[b]	18.2%	35.7%	28.6%	30.4	31.9	33.1
Differentiated instruction[g]	0.644	0.784[a]	63.6%	57.1%	42.9%	34.0	33.7	36.6
Research design[f]	0.822	0.903[a]	45.5%	42.9%	57.1%	24.0	28.7	30.2
Assessment[f]	0.825	0.904[a]	54.5%	50.0%	50.0%	11.5	22.5	20.2
Standards[g]	0.786	0.880[b]	36.4%	14.3%	21.4%	12.7	28.2	27.0
Analysis of results[f]	0.817	0.899[a]	36.4%	28.6%	28.6%	26.5	46.5	60.0
Advanced curriculum design[g]	0.797	0.887[a]	36.4%	21.4%	14.3%	34.8	21.0[e]	31.2[e]

Note. [a]Significant ($p < 0.000$). [b]Significant ($p = 0.001$). [c]Differences between Groups A and B are not significant ($p = 0.147$). [d]Differences between Groups A and C are not significant ($p = 0.083$). [e]Differences between Groups B and C are significant ($p = 0.038$). [f]Variable included in Component 2, design and analysis. [g]Variable included in Component 3, instructional planning.

Mathematics courses taken at the master's level clearly favored Groups A and C. Respectively, 6 of 11 (54.5%) and 8 of 14 (57.1%) teachers in these groups took additional mathematics courses while only 3 of 14 teachers (21.4%) in Group B did so equating to one course each. The difference in the recollection of courses between

Groups B and C was not significant ($p = 0.061$) nor were the differences in hours reported by Groups A and B ($p = 0.079$) and Groups B and C ($p = 0.051$). These data may be found in Table 19.

Table 19
Teachers' Recollection and Average Hours of Additional Mathematics Courses at the Master's Degree Level

Recalled Instruction				Average Hours of Instruction (for those taken)			
All	Group A	Group B	Group C	All	Group A	Group B	Group C
48.7%	54.5%	28.6%	64.3%	56.2	60.0	40.0	59.4

Coursework beyond master's degree was completed by 10 teachers who averaged 93.5 instructional hours each. These hours were associated with six Group A and four Group C teachers only with approximately the same average number of instructional hours. Group B teachers reported no post-master's degree courses. Table 20 presents this information.

Table 20
Teachers' Recollection and Average Hours of Coursework Beyond Master's Degree

Recalled Instruction				Average Hours of Instruction (for those taken)			
All	Group A	Group B	Group C	All	Group A	Group B	Group C
25.6%	54.6%	0.0%	28.6%	93.5	93.3	0.0	93.8

Professional Development and Training Topics

In the area of professional development and training, differentiating between school groupings was not as strongly supported by the instrument used as in other categories. The recollection of training attended, in 7 of 10 topics, was not less than 82%.

In ordinal responses, reporting the number of hours of training attended, 70% selected the maximum option, "greater than 10 hours." It was obvious from ensuing discussions with participant teachers that training, in any form, was held multiple times a week if not daily. Almost every discussion among teachers concerning the use of formative testing to inform instruction had the potential of becoming a prolonged training session with an instructor, an opinion leader, or specific goals driving the discussion. Of particular note were the comments that suggested those teachers with experience in special education frequently provided the needed expertise sought for instructional planning and revision or differentiating instruction for individual students.

Group C teachers recalled only 64.3% participation in evaluation, a low for all training topics except action research. For the evaluation topic, the mode of training attended was "None" representing 35.7% of Group C teachers. Action research, while offered by the district and supported by the literature, was the one topic not well subscribed. An average of 30.8% of all teachers had participated in action research training with 36.4% of Group A and 42.9% of Group B teachers recalling participation. Group C schools recalled 14.3% participation. The differences in recollection between Group B and C teachers were not significant ($p = 0.102$). For hours of training reported, the "None" response was the mode for 85.7% of Group C teachers. In hours, 63.4% and 57.1%, respectively, for Group A and B teachers reported no training. These data are detailed in Table 21 providing percentages of training recalled and modes of training ordinals with the percentage of teachers reporting each.

Training in test results analysis was recalled by 100%, 85.7%, and 92.9%, respectively, of Group A, B, and C teachers. However, 54.5% of Group A and 35.7% of

Table 21
Teachers' Recollection and Ordinal Hours of Training in Selected Topics with Correlation and Reliability Data

Training Topics	Pearson's *r* and Cronbach's ∀ of recollection to training hours		Training Recalled (percentage)			Mode of Training (ordinal hours of training)[a] [teachers' reporting mode][b]		
	r	∀	Group A	Group B	Group C	Group A	Group B	Group C
Total Teachers			11	14	14			
Math methods	0.598	0.749[c]	90.9%	92.9%	100%	>10 [6]	>10 [6]	>10 [9]
Assessment[l]	0.571	0.728[c]	81.8%	100%	92.9%	>10 [5]	>10 [6]	>10 [7]
Testing[j]	0.645	0.784[c]	90.9%	92.9%	92.9%	>10 [7]	>10 [6]	5 to 10 [5]
Results analysis[l]	0.652	0.789[c]	100%	85.7%	92.9%	>10 [6][f,g]	>10 [5][f]	5 to 10 & > 10 [4][g]
Data-driven decision-making[l]	0.600	0.750[c]	100%	92.9%	92.9%	2 to 5 & >10 [4]	>10 [8]	>10 [7]
Action research	0.902	0.949[d]	36.4%	42.9%[c]	14.3[c]	None [7]	None [8]	None [12]
Evaluation[l]	0.803	0.890[c]	81.8%	85.7%	64.3%	>10 [5][h]	All but None [3]	None [5][h]
Differentiated instruction[l]	0.388	0.559[c]	100%	92.9%	100%	>10 [9][i,j]	2 to 5 & >10 [4][i]	>10 [5][j]
Strategies, activities, and revision	0.592	0.743[c]	81.8%	100%	100%	>10 [8]	>10 [7][k]	>10 [8][k]
Instructional planning and revision[l]	0.800	0.889[c]	81.8%	85.7%	92.9%	> 10 [6]	>10 [7]	>10 [6]

Note. [a]Ordinal choices included 0 hours (none), 2 or less hours (<2), 2 to 5 hours, 5 to 10 hours, and more than ten hours (>10). [b]Each report includes the mode of ordinals and the percentage of subjects reporting each. [c]Significant ($p < 0.000$). [d]Significant ($p = 0.001$). [e]Differences between B and C are not significant ($p = 0.102$). [f]Difference between Groups A and B are not significant ($p = 0.120$). [g]Differences between Groups A and C are not significant ($p = 0.082$). [h]Differences between Groups A and C are not significant ($p = 0.066$). [i]Differences between Groups A and B are significant ($p = 0.039$). [j]Differences between Groups A and C are not significant ($p = 0.096$). [k]Differences between Groups B and C are not significant ($p = 0.110$). [l]Variable included in Component 1

Group B teachers reported a mode of "greater than 10 hours" of training. Group C teachers' responses were bi-modally distributed between "5 to 10 hours" and "greater than 10 hours" with 28.6% reporting each. Statistically, the difference between Group A

teachers' hours and Group B teachers and those between Group A and Group C teachers' training hours were not significant ($p = 0.120$ and $p = 0.082$, respectively). A summary of these data may be found in Table 21.

Evaluation as a training topic was recalled by 76.9% of all teachers. However, only 25.6% of all teachers reported "greater than 10 hours" training. Group A was represented by 45.5% of teachers with "greater than 10 hours" training, Group B teachers responses were nearly evenly distributed in all training response options while 35.7% of Group C teachers had a response mode of "None." The differences between Group A and Group C teachers was not significant ($p = 0.066$). A report of these data may also be found in Table 21.

Training in differentiated instruction revealed some of the more notable differences in teachers' participation. While recalled training in the topic was 97.4% for all teachers, the highest for any training topic, the hours of training reported suggested a different level of interest and participation. Group A teachers (81.8%) reported a mode of "greater than 10 hours" participation while 35.7% of Group C and 28.6% of Group B teachers reported a mode of "greater than 10 hours" of training. Statistically, the difference between Groups A and B was significant ($p = 0.039$) while the difference between Groups A and C was not ($p = 0.096$). These data are listed in Table 21.

The final area of interest occurred in the topics of training in the selection and usage of strategies, activities, and exercises. Nearly all teachers (94.9%) recalled training in this topic, however, 72.7%, 50.0%, and 57.1% of teachers in Groups A, B, and C, respectively, reported a mode of "greater than 10 hours" of training. The difference between Groups B and C was not significant ($p = 0.110$). These data are also recorded in

Table 21.

The consideration of teacher expertise as expressed by those who participated as training instructors is presented in Table 22. Groups A and C appeared to have some experiential advantage though not to levels of significance.

Table 22
Recollection of Training Provided and Instruction Ordinals

Recollection of Training Provided				Instruction Provided Ordinals				
All	Group A	Group B	Group C	Ordinals (hours)	All	Group A	Group B	Group C
33.3%	36.4%	28.6%	35.7%	0 to 2	1	0	1	0
				2 to 5	4	0	2	2
				5 to 10	2	1	0	1
				> 10	6	3	1	2

Research Question Findings

Using the variables selected and data collected, principal component analysis (PCA), a form of factor analysis, was conducted on a sample of subjects with an *n*-size of 46. Multiple criteria were used to guide the analysis. The Kaiser-Guttman's Criterion, i.e., eigenvalues greater than 1.00 (Stevens, 2002), was used to determine cut-off points for numbers of components. Three components met this criterion. A critical value for factor loadings was determined to be 0.800 based on Cliff and Hamburger's work in which doubling "critical values for correlation coefficient at $\alpha = 0.01$ for a two-tailed test" (1967, p. 394) and regressed for a sample size of 46. Five of eight variables of component one met this criterion. Variables were deleted from PCA consideration based on Field's (2005) suggestion to remove variables dependent on their Measure of Sampling Adequacy (MSA). To remain conservative, Field suggested retaining only those variables

with Kaiser-Myers-Olkin (KMO) statistics greater than 0.7. All variables remaining in the principal component analysis exceeded this criterion.

The final principal component analysis resulted in 15 variables comprising three components which accounted for 67.63% of the variance. While this variance figure did not achieve the suggested 75% (Stevens, 2002), it exceeded the 50% recommended by Tabachnick and Fidell (2001, as cited by Meyers, Gamst, & Guarino, 2006), and appeared to be a reasonable solution. An Oblimin rotation with Kaiser Normalization was used as, at their roots, the variables are related on a theoretical dimension. Data from that rotation are presented in Table 23. Using a recommendation of Field (2005), the analysis was also conducted with a Varimax rotation and Kaiser Normalization which did not achieve a mirrored component transformation matrix. A mirrored matrix, Field suggests,

Table 23
Summary of Principal Component Analysis with Oblimin Rotation

		Component Loading			
Variable	Item	1	2	3	Communality
O04	Training in analysis of test results	0.885	0.703	-0.009	0.784
O10	Training in instructional planning and revision	0.874	0.035	0.198	0.731
O03	Training in testing	0.874	-0.014	-0.077	0.795
O02	Training in assessment	0.861	0.097	0.156	0.706
O05	Training in data-driven instruction	0.855	0.213	0.008	0.755
O07	Training in evaluation	0.662	-0.356	0.433	0.782
O08	Training in differentiated instruction	0.553	-0.340	-0.422	0.611
Q11	Graduate course in research design	0.085	0.779	0.025	0.594
Q08	Graduate course in testing	0.046	0.759	-0.107	0.642
Q12	Graduate course in assessment	-0.165	0.685	-0.336	0.755
Q14	Graduate course in testing analysis	0.167	0.656	-0.135	0.533
B14	Graduate course in standards	-0.080	0.104	-0.749	0.608
B11	Graduate course in differentiated instruction	-0.021	0.135	-0.727	0.607
B16	Graduate course in advanced curriculum	0.038	0.342	-0.600	0.623
B07	Graduate course in evaluation	-0.040	0.445	-0.521	0.620
Eigenvalues		5.015	4.009	1.121	Total Variance
% of Variance		33.430	26.725	7.474	67.629%
Coefficient Alpha		0.916	0.805	0.804	

is expected if components and variables were unrelated.

Criteria that would suggest that the analysis was valid included a final Kaiser-Myers-Olkin measuring of sampling adequacy of 0.795, or "good" in Kaiser's assessment (0.7 to 0.8 is "good" adequacy, 0.8 to 0.9 is considered "excellent"). Bartlett's Test of Sphericity was significant at $p < 0.000$ suggesting an equality of variances across and covariances between conditions. Regarding residuals, there were 50 (47.0%) non-redundant residuals. Field suggests that more than 50% of non-redundant residuals are reasons for concern (2005). Cronbach's α was used to examine the reliability of the three components revealed resulting in figures of 0.916, 0.805, and 0.804. Each was significant ($p < 0.000$). In summary, as exploratory principal component analysis findings, the extracted components should have value in examining the research questions (Field, 2005).

Research Question 1

As to Research Question 1 (RQ1), "Are the focus and selected content of teacher preparation programs and graduate education variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as directed and measured by quarterly assessment results?", principal component analysis revealed two components that would suggest a relationship to a teacher's capacity to use formative testing to inform instruction. Comprised of graduate level education courses, these components and the variables that they represent make logical sense as likely contributors.

The first of two components supporting the impact of education on a teacher's use of formative testing is based on the numbers of hours in graduate courses involving research design, testing, assessment, and testing analysis. The variables comprising this component

had loading factors of 0.779 (graduate course – research design), 0.759 (graduate course – testing), 0.685 (graduate course – assessment), and 0.656 (graduate course – testing analysis), respectively. While these do not meet the critical value of 0.800 for a sample size of 46, Guadagnoli and Velicer (1988) suggest that factors or components with four loadings greater than 0.600 are reliable regardless of sample size. This suggests that the component should be useful. The component contributed 26.725 % of variance. When tested for reliability, Cronbach's ∀ was determined to be 0.805 which was significant ($p <$ 0.000). The variables comprising this component are associated with better appreciating the goals and design of research or process, albeit testing instead of research, and analyses of appropriate testing. It was tentatively named program design and analysis.

The second component supporting Research Question One was based on dichotomous data (yes-no) representing recalled participation in graduate courses concerning standards, differentiated instruction, advanced curriculum design, and evaluation. While the weakest of the three components revealed, its variables displayed loading factors of -0.749 (graduate course – standards), -0.727 (graduate course – differentiated instruction), -0.600 (graduate course – advanced curriculum), and -0.521 (graduate course – evaluation) which nearly meets the Guadagnoli and Velicer (1988) suggestion. In light of the one variable, evaluation, with a loading factor less than an absolute value of 0.600, Naik (personal communication, July 10, 2006), suggested, owing to the small sample size of 46, retaining the component for further study, re-examination, and confirmation with a larger sample during a future study. The fact that all loading factors are negative does not cast doubt on the value of the component. The condition, referred to as reverse phrasing, represents a condition in which the variables may be perceived as opposite to the remainder

of variables (Field, 2005). It is also noted that combinations of positive and negative loading factors can place limitations on the use of reliability analyses, such as Cronbach's ∀, as negative and positive contributors tend to cancel one another and induce error and produce inconclusive reliability figures. As all loading factors were negative, the ambiguity is non-existent. The significant ($p < 0.000$) Cronbach's ∀ of 0.804 is of value. The second educational component, tentatively named instructional planning, had an eigenvalue of 1.121 and accounted for 7.474% of the variance. This component appeared to be tentatively associated with understanding the standards-based necessity of instruction and the need to differentiate for individual students. In summary, the second and third components extracted during principal component analysis were based on graduate education and tentatively identified as program design and analyses and instructional planning, respectively. The analysis suggests that these components are of value in supporting a teacher's capacity to use formative assessment to inform instruction.

Research Question Two

Regarding Research Question Two, "Are selected contents of professional development or in-service training programs variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as both directed and measured by quarterly assessment results?", principal component analysis revealed a component comprised of professional development and training topics that would appear to strongly contribute to a teacher's capacity to use formative testing to inform instruction. The component included seven variables derived from ordinal representations of hours in these training topics. These variables had loading factors of 0.885 (topic - analysis of test results), 0.874 (topic - instructional planning and revision), 0.874 (topic - testing), 0.861

(topic - assessment), 0.855 (topic - data-driven instruction), 0.662 (topic - evaluation), and 0.553 (topic - differentiated instruction). Five of these loading factors were in excess of the critical value of 0.800, while six were greater than 0.6. This component appears to support the value of professional development and training as a positive contributor to a teacher's capacity to use formative testing to inform instruction.

Summary

In this chapter, the population and sample of teachers responding to the study were briefly discussed including the events and decisions that shaped and resulted in the final sample from which data were collected. The stratification scheme used for the preliminary review of data was highlighted, resulting in Groups A, B, and C representing schools whose students were high, median, and low achievers. The instrument and data coding used were presented discussing the nature of the data and their importance to the analyses. Protocol analyses' results were provided for the instrument used (Appendix A) including Cronbach α's for variable pairs representing dichotomous and quantitative representations of courses or topics experienced. The impact of self-reported and possibly time-biased responses was considered. The emergence of an educational foci ordinal variable from nominal identifications of fields of study was introduced and defined. These items were presented in preparation for preliminary analyses.

Preliminary analyses of the data were provided for the entirety of teacher histories collected in order to understand the components revealed. Results of school rankings using the Colorado Student Assessment Program model were provided. The impact of class size was examined. It was suggested that class size may not have been a factor in achievement as Group A had the largest average class size (20.7 compared to Group B's

18.3 and Group C's 19.3). Teaching experience was examined, overall and in third grade, noting that Group A schools, on average, had nearly a seven year advantage over Group B and C schools in overall experience.

Teacher's undergraduate educational programs' foci were identified and placed in one of five groupings including special education, education, psychology/sociology, content, and other. It was observed that Group A and B schools had the advantage of more special education-oriented teachers, while Group C schools had the greatest number of non-educational track undergraduate experiences. In graduate education, the nuance of master's degree as part of an initial licensure effort and master's degree as classic graduate education was suggested and reconstructed from the data. All graduate degrees were in the field of education and nearly equally distributed in all groups although degrees associated with initial licensure were more prevalent in Group C schools.

Likert-scaled items involving a teacher's personal *like* or preference for mathematics, belief in the formative testing program as practiced, and each teacher's evaluation of his/her mathematics specialist's assistance were discussed. It was noted that Group A and C teachers were largely positive in their *like* of mathematics. The presence of more neutral or negative *like* for mathematics was noted in Group B. Belief in the formative testing program was more strongly expressed by Group A than by Groups B or C. It was noted that no teachers in Group B strongly believed in the program while Group C teachers were evenly divided between positive beliefs and neutral or negative beliefs. Regarding evaluation of each school's mathematics specialists, it was interesting to note that Group A teachers had strong opinions, both negative and positive, Group B teachers responses ramped toward the negative, and Group C seemed largely satisfied. It was also

noted that the district's mathematics curriculum coordinator's evaluations of mathematics specialists had a low correlation, overall ($r = 0.254$), and a negative correlation with Group B teachers ($r = -0.655$). Correlations with Group A and B teachers were moderate ($r = 0.608$ and $r = 0.567$, respectively). The negative correlation is an item of interest.

Regarding selected topics in undergraduate education, additional instruction in mathematics appeared to favor, in order, Groups A, B, and C. Group A schools also had an advantage in exposure to assessment and testing. Less than one-third of all teachers recalled instruction in item writing and, for those that did, less than four hours of instruction were noted. Instruction in data-driven instruction was recalled by about one-fourth of teachers but, for those that did, no fewer than 25 hours were reported. This topic favored Group C teachers.

Graduate course participation produced greater numbers of hours of instruction recalled and hours reported. It was noted that there was a disparity in recollection and numbers of course hours reported. Courses in differentiated instruction and test item construction produced the greatest number of hours reported. Participation in differentiated instruction courses was about equivalent for all groups with an average of 35 hours of instruction per participant. With the exception of advanced curriculum design as a course, there were no significant differences in participation among the three groups. The one exception produced a significant difference between Group B and C teachers. Considering mathematics classes taken at the graduate level, Groups A and C had an advantage over Group B. Regarding courses taken beyond a master's degree, again Groups A and C had an advantage over Group B as no teacher in Group B reported taking such courses.

Professional development, in-service, and capacity development training was presented with the observation that there was little to differentiate among groupings of teachers. Except for training in action research, the lowest recalled participation in any topic was Group C with only two-thirds of teachers recalling training. Action research was the least subscribed training with an average of nearly 70% reporting no training. Training in differentiated instruction was the most recalled training as all but one teacher reported having attended. Mathematics methods and training in strategies, activities, and exercises were, by hours and participation levels reported, the most subscribed training. Training, in terms of hours attended, favored Group A in 7 of 10 topics.

Research question findings were discussed. The principal component analysis resulted in three components accounting for 67.6% of the variance. The criteria used to guide the analysis were briefly discussed. These seemed to confirm the components revealed were reliable. Cronbach's α's for these three components was reported as significant ($p < 0.000$) with none being less than 0.804. The three components were tentatively identified as professional development, program design and analyses, and instructional planning.

Research Question One dealt with the impact of education on a teacher's capacity to use formative testing to inform instruction. As graduate education course variables were found to contribute to two of the three components identified, these components may suggest that education, and in particular graduate education, is an important contributor to the formative testing to inform instruction process. Research Question Two addressed the impact of professional development on a teacher's capacity to use formative testing to inform instruction. This component is the strongest component and

was comprised of seven training variables. These data suggest that professional development is a potentially strong contributor to the capacity to use formative testing to inform instruction.

Chapter V will provide a summary of the findings provided. The value of the preliminary analyses will be clarified as well as the conclusions they support in the effort to understand the components revealed. Conclusions will be stated summarizing the findings drawn from data supporting the research questions. Recommendations for future studies regarding the research questions or related topics will be made.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This chapter will begin with a restatement of the problem, presentation of the research questions, and the highlights of the study's background, significance, setting, limitations, and assumptions. A synopsis of the literature's salient points regarding this study will be followed by a brief review of the methodology employed, the sample obtained, the findings reported, and the results of the principal component analysis conducted. Conclusions will then be drawn regarding each of the research questions. Each outcome will be briefly discussed. The chapter will conclude with recommendations regarding the review of current teacher education and training policies and practices as well as those regarding future research into the development of teachers' capacities involving formative assessment and testing.

Summary

The goal of the study was to examine teachers' education and training histories that could possibly contribute to their capacities to use formative assessment to inform instruction as measured by their students' achievement in third grade mathematics. Two research questions were posed. The first sought to determine if participation in select content of teacher preparation programs and graduate education were contributory to their capacities to convert formative testing results into informed instruction. An examination of potential contributions made by select undergraduate, graduate, and continuing education topics or courses was used to achieve this goal. The second question sought to determine if professional development and in-service training programs in select topics were contributory to a teacher's capacity to inform instruction. Study of potential professional development and

training contributions included those provided within school and district venues. The study's approach treated participation and time spent in the selected instructional sessions of both types as variables whose variations likely reflected aspects of the capacity to use formative testing to cyclically inform instruction thereby improving learner achievement.

The background of the study focused on standards-based educational reform and its preliminary emphasis on and use of summative assessment in the form of high-stakes, state-level testing to promote academic success and provide for public accountability. These high-stakes tests were not viewed as helpful to teachers or students (Glaser & Silver, 1994). Many issues associated with standards and assessments were examined by researchers who suggested that there were aspects more critical to the approach's success than summative testing. The way assessment results were to be used (Darling-Hammond, 1994a), the opportunities to create new forms of assessment (Baker, 1994), and more formative uses of assessment (Black & Wiliam, 1998a, 1998b) were highlighted. Warnings that educator perceptions and uses of assessment had to change were also evident suggesting that practice had to move from the *assessment of learning* to *assessment for learning*. This difficult shift in paradigm was considered a challenge that had to be met (Koretz et al., 1996). New, creative, fair-minded, and equitable ways of using assessments to inform instruction were required. In accomplishing these tasks, teachers would need assistance (Darling-Hammond, 1994a).

The significance of the study was stated to be a contribution to better understanding formative testing and informed instruction processes and teachers' capacities to employ them. The first of three possible contributions was an examination of processes inside the "black box" of assessment (Black and Wiliam, 1998b), e.g., teachers' capacities to guide instruction with test results. A better understanding of the relationship between educational and training

contributors to a teacher's capabilities to use assessment was the second. The third, and possibly the most tangible, was a determination of the relative impact of education (undergraduate, pre-licensure, graduate, and continuing) and professional development on teachers' capacities to successfully guide instruction with testing results.

The study's setting was established to be a city school district in Virginia. Students were characterized as largely economically disadvantaged, displaying achievement gaps along ethnic lines. In an effort to raise achievement and close gaps, the district opted to employ a quarterly, formative, content assessment program that encouraged teachers to guide subsequent instruction based on testing results (Strauss, personal communication, July, 2003).

Limitations of the study included concerns for non-teacher influences on test results, test analyses, and instructional revision. Caution was expressed regarding the use, depth, and accuracy of self-reported data. Variability in the venues, topics, and instruction reported were cited as sources of variation potentially compromising the value of data collected. School-level instructional and testing models and their potential effects on a study of teachers' capacities were expressed as potentially confounding forces. Finally, the potential non-generalizeability of the study's results – based on the use of a single content area, single instructional model, and a specific periodicity of formative testing – was stated.

The potential impact of elements not measured such as instructional effectiveness, school organization, and socioeconomics were considered as assumptions. It was assumed that a school's instructional model influenced teacher preparedness, e.g., preparation to teach four content areas versus two, making a difference in student outcomes. School-wide testing more frequent than the district's testing program was assumed to be a potential source of confounding effects. The merits of district assessments were assumed to have the same

effects district-wide. District assessments were assumed to be low-stakes testing hence not subject to unethical practices by teachers or schools in their preparation for or administration. And, finally, it was assumed that access to required data would be granted.

Review of the literature began with a history of assessment, the dominance of summative forms, and near constant efforts to reform public education following World War II. These topics ushered in state standards and achievement testing as means to achieve reform (Bedwell, 2004; Engelhart, 1950; Glaser & Silver, 1994; Haladyna, 2004; Hamilton et al., 2003; Linn, 1987, 2000; Mislevy et al., 2001). The inter-reliant nature of standards and assessments was discussed (Yoon & Resnick, 1998). High-stakes state testing was noted to boost the continued domination of summative assessment. States' attraction to high-stakes testing programs was explained as relatively inexpensive ways to mandate reform while providing for public accountability (Linn, 2000). Researchers' examinations of early state assessment programs highlighted the pivotal importance of teachers' behaviors in achieving reform goals. Necessary behaviors were seen as changes difficult to achieve, requiring professional development, and dependent on each teacher's shift in perspective concerning assessment (Koretz et al., 1996). Congruent with earlier projections of such needs (Darling-Hammond, 1994a), the literature of the late 1990s shifted its foci from summative to formative assessment.

The shift in research foci from the assessment of learning to assessment for learning was important (Assessment Reform Group, 1999; Black & Wiliam, 1998a, 1998b). Formative assessment was seen as being critical to achieving the goals of the standards and assessments movement (Black & Wiliam, 1998b). Formative assessment research indicated that is was a powerful method in closing the gap between standards-

based requirements and achievement levels with the added benefit of having greater impact on lower achieving students (Black & Wiliam, 1998a). The cyclic nature of formative assessment was discussed by many researchers as was the importance of appropriate and timely feedback (Black & Wiliam, 1998a, 1998b; Costa, 1993; Crocker & Algina, 1986; Ebel & Frisbie, 1986; Glaser, 1962; Skinner, 1963; Gronlund & Linn, 1990; Ramaprasad, 1983; Ravitz, 2002; Sadler, 1989; Schafer, 1991, 1993; Sleight, 1993; Stiggins, 1991, 2002; Taylor, 1994; Wiggins, 1994; Wiliam & Black, 1996; Vygotsky, 1962). Formative assessment feedback processes were also seen as critical steps in developing students' capacities to self-assess (Black & Wiliam, 1998a).

Despite the consensus of researchers' opinions regarding best practices and formative assessment, summative assessments' dominance continued (Black et al., 2004; McMillan, 2003; McMillan et al., 2002; McNair, et al., 2003). Causes, research suggested, included a continued lack of educational and training preparation in the use of formative methods (Black, 2000; Johnston & Lawrence, 2005; Marzano, 2003; Stiggins, 2001a, 2001b, 2002, 2005). Failure to achieve the needed cultural paradigm shift in education and absences of actions necessary to produce such changes were also contributory to enduring summative practices (Darling-Hammond, 1994a; Koretz et al., 1996; McMillan, 2003; Shepard, 2000; Stiggins, 2001a). Research discussing teacher preparation and professional development for assessment suggested what content was missing from the substance of instructional programs (Black & Wiliam, 1998a, 1998b; Darling-Hammond, 1994a, 1994b; Gronlund & Linn, 1990; Marzano, 2003).

Content that teachers should know to use formative assessment was reviewed. Black & Wiliam's (1998b) observation that teachers generally did not have time to weave

principles into practices but needed practical instruction in their application highlighted the form of instruction necessary. McMillan (2003) discussed reasons for non-adoption of new ways. The impact of beliefs on all things educational, whether concerning students' abilities to learn or the worth of assessment for learning practices, was insightful. Many researchers' observations about beliefs could be generalized as defining the possible and the unlikely, filters that determined what was achievable and how course materials, programs, and processes would be assimilated by individual students and teachers.

Testing was reviewed as it pertained to the use of results to drive instruction (Airasian, 1998; Bracey, 1987; Cizek, 1993; Frederiksen, 1994; Haertel & Herman, 2005; Herman, 1997; Madaus, 1988; Popham, 1987; Shepard, 1991; Tyler, 1949). The interactions of standards (e.g., SOLs), curriculum guides, and pacing in the sequencing and timing of instruction and local testing were explored as was the impact of pacing disconnects on the validity of results. Standards and testing were discussed (Reigeluth, 1997) as was the need for both to be comprehensive and unambiguous (Anderson & Krathwohl, 2001; Berk, 1980; Marzano, 2000; Schmoker & Marzano, 1999). The impact of cognitive tasking was examined suggesting its importance to instruction, learning, testing, and results (Anderson & Krathwohl, 2001; Baker, 2004; Bloom et al., 1956; Marzano, 1992, 2001). The need for instruction and testing results to be differentiated was addressed (Barr & Dreeben, 1997; Council of Great City Schools, 2003, 2004; Wiliam & Black, 1996) as was the virtue of conducting assessments district-wide as opposed to single classrooms (Shepard, 2001).

The importance of integrating instruction and formative testing was reviewed. The contention that the culture of learning was established during instruction was voiced

(Black & Wiliam, 1998b; Shepard, 2001). Almond, Steinberg, and Mislevy's Four Principle Processes in the Assessment Cycle (2003) was introduced as a model for the instruction - assessment cycle. It included activity selection, presentation, response, and summary scoring. The skills needed for successful instruction were reiterated including the impact of beliefs (Cohen & Ball, 1999; Marzano, 2003). The depth of teachers' content knowledge was questioned suggesting that it might not always be sufficient to provide the detailed feedback needed (Cohen & Ball, 1999; Delandshere & Jones, 1999; Shepard, 2003). As classic elementary teachers are generalists, this observation was salient. The impact of planning on instruction and assessment was discussed with strong arguments suggesting that the two were often "decoupled systems" (Glaser & Silver, 1993, 1994). It was advocated that the line between instruction and assessment be blurred, where feedback was constant and timely (Anderson & Krathwohl, 2001; Brookhart, 2003; Koretz et al., 2006). Standards, biases, equal access and opportunities to learn, appropriate forms of instruction, reliability and consistency of testing (Buckendahl et al., 2002), clear, non-clumped standards in instruction, testing, and feedback (Baker, 2002), planning to construct levels (Popham, 2003), and appropriate levels of cognitive challenge (Anderson & Krathwohl, 2001; Haladyna, 2004) were also considered planning issues. The ethics of instruction and assessment were discussed (Haladyna, Nolen, & Hass, 1991) as was a proper emphasis on teaching, working harder, and working more efficiently while avoiding inappropriate preparation for assessment (Koretz et al., 2001).

Regarding assessment and formative testing, Stiggins stated that assessment had been the "victim of gross neglect" (2001, p. 10) and suggested that it was necessary to retrain 2.5 million teachers and administrators in its proper usage (1999). McMillan

(2003) suggested teachers' assessment practices were highly individualized, seldom conformed to best practices, and were usually learned from their peers. Lewis (2001) and Marzano (2003) highlighted political and technological influences that often compelled districts to seek inappropriate solutions to assessment and testing needs.

Knowledge of testing and its many elements were presented as content to which a teacher should be introduced (Berk, 1980; Cizek, Crocker & Algina, 1986; Cunningham, 1986; Gronlund, 1998, 2006; Haladyna, 2004; Mislevy et al., 2002; Popham, 1971, 1999, 2004) including the potential for errors and procedures to reduce them (Bedwell, 2004; Koretz et al., 1993; Stecher & Klein, 1997; Stecher et al., 1997; Stiggins, 2001; Wong & McGraw, 1999). Technology, often seen as the solution to assessment problems, was presented as a potential peril (Lewis, 2005; Ravitz, 2002). Regarding test results analyses, emphasis was given to understanding not computation (McMillan, 2000), systematic analysis (Shepard, 2000b), and the need for process transparency (Black & Wiliam, 1998b; McMillan, 2003; Yoon & Resnick, 1998). The use of analytic results to guide future, possibly differentiated, instruction was advocated with results specified for each learner and substandard (Athanases & Achinstein, 2003; Brimijoin, Marquiesse, & Tomlinson, 2003; Brown & Capp, 2003; Council of Great City Schools, 2003, 2004a, 2004b; Sachs, 2004; Yorke, 2003). A teacher's need to be practiced enough to accomplish these processes in a timely fashion was reiterated (Black & Wiliam, 1998a; Brown & Capp, 2003; Gronlund, 1998; Marzano, 2003).

Pre-licensure education was examined in light of expectations provided by several cooperating national organizations. Coordinated through the National Council for the Accreditation of Teacher Education (NCATE), requirements established by the National

Council of Teachers of Mathematics (NCTM), the National Association for the Education of Younger Children (NAEYC), the Association for Childhood Education International, and the Interstate New Teachers Assessment and Support Consortium (INTASC), as well as the Commonwealth of Virginia's licensure requirements, were considered. The combined requirements and impact on the preparedness of teachers were explored as was the failure to add assessment standards to licensure requirements (Stiggins, 2002).

Graduate education was first considered as that intended to complete licensure requirements beyond typical undergraduate programs. The second form was identified as classic mastery of a discipline and development of research skills (Eisenberg, 1999; Heathcott, 2005; Mislevy 1996a). Doheny (2002) suggested the latter was about the examination/re-examination of the body of research, theories, and practices. These efforts were seen to promote change in teachers' insights thereby enhancing student learning (Haladyna, 2004; Johnson & Button, 2000; Sax & Fisher, 2001). Graduate programs in education were examined and considered refinements of educational topics and extension into advanced concepts. A graduate research component, including evaluation, research design, and statistics, was highlighted. Weaknesses of graduate programs were presented. The vulnerability of programs to external forces was addressed (Heiss, 1968). Gilbert and Smith (2003) observed that graduate studies were often seen as never-ending, frustrating, time-consuming, and ever more expensive by students. Courses were often seen as being too conceptual to be immediately useful. It was Thorndike (1997) who suggested that staying abreast of one's field was something other than graduate education.

Professional development and in-service training were examined. Considered an ethical issue (Thorndike, 1997), professional growth and constant retraining were seen as

ways to eliminate obsolete practices. It was suggested that in-service training provided the greatest potential for meeting teacher's assessment needs (Stiggins & Bridgeford, 1985). Professional development was cited as being reliable, systematic, and focused on ways in which to improve existing teachers in any school or district. With time, clear objectives, resources, and appropriate instructors, any number of deficiencies could be corrected. Professional development should have the trust of the teachers involved (Borko, 1997 or 2004), be well-developed and proactive (Stecher et al., 1997), relevant to teachers' tasks, and include content knowledge as well as the arts and science of education (Stotsky, 2006). The styles and beliefs of teachers were considered fair topics (Cohen & Ball, 1999; McMillan, 2003). Leadership of such training, it was suggested, was best provided by teachers whose learners demonstrated the exemplary achievement desired (Stecher et al., 1997). If, as Black and Wiliam's (1998b) contended, that many teachers' problems were time-management issues, the training must also be time-efficient. Formative assessment training was seen to require foundations in standards, curriculum, and measurement (Cohen & Ball, 1999), be properly developed, and the product of long-term, comprehensive efforts (Stiggins, 2002; Borko, 1997). Professional development in assessment, it was summarized, must have the depth, content, and the support necessary to make the effort worthwhile. Previous efforts, noted Borko (2004), were typically fragmented, superficial, and did not account for the ways in which teachers learn. Professional development, it was summarized, properly supported, executed, and received, could make a considerable contribution to the assessment solution.

The methods and procedures proposed and used in this study are herein addressed simultaneously in order to reduce redundancy. This research, characterized as an

exploratory, *ex post facto* study, examined the educational and training histories of in-service, third grade teachers in an effort to associate their education and training experiences to their capacities to use formative assessment to inform instruction. Collecting a robust number of variables, principal component analysis was used in hope of producing a few components that might be representative of the capacity to use formative testing to guide instruction. It was suggested that components revealed might represent latent traits or elements that underlay the formative process but this is true only if corroborated by common factor analysis (Bandalos & Boehm, 2007).

To remove potential confounding factors, school instructional and testing models data were collected from district professional development and local testing coordinators. An instructional model in which one teacher taught all subjects and a testing model in which school-wide testing was done only through the district program were the foci. The goal of these limitations was reduction of potentially confounding variations. These variables are listed in Section I of Table 24. Data for variables representing the district's mathematics curriculum coordinator's evaluation of teachers' capacities to use test results to inform instruction and participant schools' mathematics specialists were collected by interview. These are listed in Section II of Table 24. The numerous variables for which data were collected from teachers and re-licensure submission records are listed in Section III of Table 24. Quarterly mathematics assessment scores and class size information were obtained from the district's assessment system. These variables are listed in Section IV of Table 24. Average student achievement scores for each teacher were used to rank schools using the Colorado Student Assessment Program (CSAP) model (Linn & Haug, 2002). Three groups were designated representing the top five,

Table 24
Research Variables

Section I - Variables collected from professional development and local testing coordinators				
School instructional model			School testing model	
Section II - Variables collected from mathematics curriculum coordinator				
Teacher's capacity to convert test results into informed instruction			School mathematics specialist evaluation	
Section III - Variables collected from teacher interview protocol				
Education				**Training**
Undergraduate or Pre-Service	Master's Degree, Graduate or Continuing			Professional Development
	Participated in graduate courses			
Bachelor's degree major	Master's degree major			
Bachelor's degree minor(s)	Additional focus area(s)			
Coursework in:	Courses in:			Topics in:
Assessment Testing Item writing Data-driven instruction	Evaluation Statistics Testing Test item construction Differentiated instruction Research design Assessment Standards Analysis of test data Advanced curriculum design Additional courses			Mathematics methods Assessment Testing Test results analysis Data-driven decision-making Action research Evaluation Differentiated instruction Strategies, activities, and exercises Instructional planning and revision
Mathematics courses beyond program requirements	Mathematics course in addition to program requirements			
	Courses beyond Master's degree			Service as professional or capacity development or in-service training instructor
	Date Master's degree awarded			
Teaching Experience				
Years teaching third grade			Total years teaching	
Opinion or Evaluations (Likert-scaled)				
Personal like/dislike for mathematics	Belief in formative assessment as practiced within the district		Assistance received from school's mathematics specialist	
Section IV - Variables collected from district assessment application				
Class size	Q1 Mathematics Scores	Q2 Mathematics Scores	Q3 Mathematics	
Section V - Variables developed as the result of analytic processes				
Strata Name	CSAP Score	CSAP Rank	Bachelor's Degree Code	Master's Degree Code

median five, and low five performing schools participating. These groups are identified as A, B, and C, respectively. Three variables were generated including group identity,

CSAP score, and CSAP rank. These variables are reported in Section V of Table 24. The stratification scheme was intended to permit examination of group differences in order to better understand any components extracted. Preliminary analysis resulted in teachers' nominal degree fields being grouped, given ordinal rankings, and added as variables. These are also listed in Section V of Table 24.

Field procedures focused on interview as the preferred form of data collection conducted within respective schools, although mailed surveys were used as a contingency. The final teachers' interview protocol was an adaptation of the Washington State Survey of Teachers (Stecher et al., 2003) shortened to comply with a district stipulation that total teacher contact was to be limited to single sessions no longer than 45 minutes. Those teachers surveyed and re-licensure records reviewed both used the interview protocol.

Prior to conducting principal component analysis, the data were statistically examined to determine instrument reliability and to generate correlations, descriptive statistics, and group differences. These analyses aided in determining the sensibility of the data collected and provided a better understanding of components that were extracted. Principal component analysis was guided by *a priori* analytic decision-making criteria.

Response to the study included 46 of 61 teachers (75.4%) in 17 schools. The original stratification scheme of Groups A, B, and C, with minor adjustments, was achieved. The distribution of schools and the 39 teachers considered in the preliminary analysis of stratified data, with group identifiers and CSAP model rankings amongst all district schools, are displayed in Table 25. Data for all 46 teachers were used in the principal component analysis.

When responses to the interview protocol were analyzed, 26 of 28 paired items

Table 25
Final Stratified Teacher Distribution

CSAP Rank[a]	Group	Participating Teacher Count
1	A	2
2	A	2
6	A	2
9	A	2
10	A	3
13	B	4
14	B	4
15	B	1
17	B	3
19	B	2
23	C	4
24	C	1
26	C	4
27	C	3
34	C	2

Note. [a]CSAP ranks are based on the 35 elementary schools within the district.

had Cronbach's ∀s greater than 0.7, a value demonstrating acceptable internal consistency (Nunnally, 1978). Preliminary analyses of the data revealed several potential contributors to differences in student performance. While class size differences were significant between Groups A and B ($p = 0.035$), Group A classes averaged 20.73 students with Groups B (18.29) and C (19.29) averaging less. Although not significant (A – B, $p = 0.072$ and A – C, $p = 0.135$), Group A teachers were more experienced (17.55 years) than Groups B (10.50) and C (10.86). Undergraduate foci on education, psychology, and sociology favored Groups A (72.8%) and B (78.5%) while 42.9% of Group C teachers majored in other areas. Positive belief in the formative assessment as practiced favored Group A (81.8%) teachers with Groups A and B differences being significant ($p = 0.040$). Evaluation of mathematics specialists found Group A teachers to

have a bimodal distribution, Group B teachers' (57.1%) being negative disposed, and Group C teachers apparently satisfied. Differences between Groups B and C were significant ($p < 0.000$). The curriculum coordinator's and teachers' evaluations of mathematics specialists had an overall Pearson's r of 0.254. Coordinator – teacher correlations for Groups A ($r = 0.608$) and C ($r = 0.567$) were moderate while correlation with Group B was negative ($r = -0.656$). Analysis of participation in undergraduate and graduate education topics revealed two significant differences; one between Groups A and C in undergraduate assessment ($p = 0.045$) and the other between Groups C and B in graduate advanced curriculum design ($p = 0.038$). Analyses of professional development and training revealed that in 7 of 10 topics, 82% or more of teachers reported participation while 70% of those participated in "greater than 10 hours" of training. Of the seven topics with high participation, only differentiated instruction produced significance as 81.8% of Group A teachers reported "greater than 10 hours" of training while only 35.7% of Group C and 28.6% of Group B did so. Differences between Groups A and B were significant ($p = 0.039$).

Principal component analysis extracted three components with eigenvalues greater than 1.00. These components accounted for 67.63% of variance. Producing a Kaiser-Myers-Olkin measure of sampling adequacy of 0.795, the analysis was considered to have "good" validity. Cronbach ∀ measures of reliability for the three components extracted were 0.916, 0.805, and 0.804, respectively with all being significant ($p < 0.000$). The strongest component (eigenvalue = 5.015) was comprised of seven training variables including analysis of test results (loading factor = 0.885), instructional planning and revision (0.874), testing (0.874), assessment (0.861), data-driven instruction (0.855), evaluation

(0.662), and differentiated instruction (0.553). The second component (eigenvalue = 4.009) represented graduate education course hours in research design (0.779), testing (0.759), assessment (0.685), and testing analysis (0.656). The final component (eigenvalue = 1.121) represented graduate education participation in standards (-0.749), differentiated instruction (-0.727), advanced curriculum (-0.600), and evaluation (-0.521).

Conclusions

The exploration of teachers' capacities to use formative testing to inform instruction resulted in 15 of 84 variables producing three principal components. The component associated with professional development and training accounted for 33.4% of variance while the second and third components, associated with graduate education, combined to account for 34.2% of variance. The extracted components were used to address the two research questions.

Research Question 1

Research Question 1 (RQ1) was, "Are the focus and selected content of teacher preparation programs and graduate education variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as directed and measured by quarterly assessment results?" The question examined participation in a select list of educational topics which, as variables, represented preparatory (e.g., undergraduate or pre-licensure graduate), graduate, and continuing education. Principal component analysis extracted two components composed of graduate education variables suggesting that participation in the associated graduate education courses does contribute to the capacity of interest.

The first of these components, based on the number of instructional hours spent in

graduate courses concerning research design, testing, assessment, and testing analysis, appeared to be valid as all loading factors (0.779, 0.759, 0.685, and 0.656, respectively) were greater than 0.600 (Guadagnoli & Velicer, 1988). Identified as program design and analysis (i.e., understanding assessment program design, goals, and processes as well as the capacity to accomplish associated analyses), this component's courses likely produced an understanding of concepts, designs, and goals useful when working with formative testing programs, the interrelationship of associated processes, and the analytic acumen sufficient to extract meaningful insights from testing results. The understanding and skills contributed, it was suggested, were representative of formative assessment issues discussed by Black and Wiliam (1998a, b), McMillan (2003), Shepard (2000), and Stiggins (2002), usage of results suggested by Darling-Hammond (1994a, b), and achievement of shifts in perspective suggested by Koretz et al. (1994). The preliminary analysis of variables representing courses in assessment and analysis favored Group A teachers while variables representing courses in testing and research design favored Group C teachers. Group differences involving these variables were not significant.

The second component, based on teachers' recollections of participation in graduate courses concerning standards, differentiated instruction, advanced curriculum design, and evaluation, is tentatively valid as the absolute values of a majority of loading factors (- 0.749, - 0.727, - 0.600, and - 0.521, respectively) were greater than or equal to 0.600. As all loading factors are negative, a condition referred to as reverse phrasing (Field, 2005), and the absolute value of the one loading factor (- 0.521) was less than 0.600 thus not compliant with Guadagnoli and Velicer's criteria (1988), the component is retained for further examination and retest with a larger sample (Naik, personal communication, July 2006). Identified as

instructional planning, the component's courses appear to be associated with understanding the influences of standards and differentiation on curriculum design, development of instruction based on testing results, and evaluation of the process at large. This component suggests that the simultaneous planning of instruction and assessment blends the processes producing seamless efforts. These efforts account for the needs of specific students while creating an environment in which constant feedback is given to and expected by students regardless of the form such feedback takes (Almond, Steinbert, & Mislevy, 2003; Anderson & Krathwohl, 2001; Black & Wiliam, 1998a, b; Brookhart, 2003; Glaser, 1962; Glaser & Silver, 1993, 1994; Gronlund, 1998; Koretz, Mitchell, Barron, & Keith, 1996; Linn, 1998; Pellegrino, Baxter, & Glaser, 1999; Wiggins, 1994). Regarding preliminary analyses, all associated variables favored Group A teachers while results for Groups B and C were ambiguous. Participation in evaluation courses produced significant differences between Groups A and B ($p = 0.021$) and Groups A and C ($p = 0.050$).

Summing the impact of both components drawn from graduate education, it is suggested that affirmation of Research Question 1 is supported. Graduate education, in selected topics, appears to create cohesive components that contribute to teachers' capacities to use formative assessment to inform instruction in third grade mathematics. Preliminary analyses of the data found that six of the eight associated variables favored Group A teachers thereby highlighting the strength of their value. Undergraduate education did not appear to influence the capacity of interest. This observation may be inconclusive as poor study design or poor recollection on the part of participants – a stated limitation – could have confounded the results. However, this likely conclusion is consistent with the literature which suggests that preparation programs are not yet sufficient to prepare teachers for standards and

formative assessment environments. While 89.7% of teachers had participated in graduate courses, 47.7% did not hold master's degrees, a possible insight into the synergistic effects of well-designed programs. It seems reasonable to suggest that those without advanced degrees might benefit from participating in specific graduate courses if not an entire program. Whether development of capacities to use formative assessment to guide instruction were part of pre-licensure graduate degrees as differentiated from advanced studies graduate programs is unknown. The data collected were insufficiently defined to support such an observation. One final observation was made regarding additional mathematics education. Education in mathematics was not associated with either component. It may be appropriate to infer that the capacity of interest has merit in any or all content areas.

Research Question 2

Research Question 2 (RQ2) was, "Are selected contents of professional development or in-service training programs variables contributory to a teacher's capacity to convert formative testing results into informed instruction in third grade mathematics as both directed and measured by quarterly assessment results?" This question examined a select list of training topics, similar to those in a teacher's education, in order to determine their contribution to a teacher's capacity to use formative testing to inform instruction. The component extracted consisted of seven professional development and training variables representing analysis of test results (loading factor = 0.885), instructional planning and revision (0.874), testing (0.874), assessment (0.861), data-driven instruction (0.855), evaluation (0.662), and differentiated instruction (0.553). Five loading factors were greater than the calculated critical value of 0.800 while six were greater than 0.600 (Guadagnoli & Velicer, 1988). The component, identified as professional development, addresses all phases

of the process of informing instruction based on testing results. Content of the training identified supported stages of the process of using results derived from district-wide tests, practical understanding of testing programs, practical usage of data-driven and differentiated instruction principles, and the capacity to incorporate insights into ensuing instructional planning. This conclusion appears consistent with the literature regarding the power of professional development and processes associated with formative assessment to inform instruction (Borko, 2004; Borko et al., 1997; Cohen & Ball, 1999; Darling-Hammond, 1994a; Delandshere, 1996; Ebel & Frisbie, 1986; Heritage et al., 2005; Impara et al., 1993; Koretz et al., 1993; Koretz et al., 1996; McMillan, 2003; Resnick & Harwell, 2000; Simmons & Resnick, 1993; Stecher et al., 1997; Stecher et al., 1998; Stiggins, 1999, 2001, 2002; Stiggins & Bridgeford, 1985; Thorndike, 1997; Yoon & Resnick, 1998). Preliminary analyses of data for these variables found all favoring Group A. Differences between Groups A and C for variables representing training in test results analysis and evaluation were significant ($p = 0.024$ and $p = 0.006$, respectively). In six of seven variables, Group C teachers were least favored.

Based on the strength of this component, it is suggested that affirmation of Research Question 2 is supported. Professional development and training in selected topics, i.e., analysis of test results, instructional planning and revision, testing, assessment, data-driven instruction, evaluation, and differentiated instruction, clearly appear to make a substantial contribution to a teacher's capacity to use formative assessment to inform instruction. Like the observation concerning educational contributions, the absence of mathematics methods and methods-oriented variables in this component may suggest that the capacity of interest is effective for all content areas.

Recommendations

The conclusions of this study link specific topics of professional development and training and graduate education to teachers' capacities to use formative testing results to inform instruction. This research, experiences gained, and conclusions contributed to several recommendations for consideration regarding policy and practice as well as future investigations.

For Policy and Practice

The conclusions of this study suggested a number of issues for consideration in the examination of existing policy and practice. Chief among these areas for review is the apparent absence of influence on teachers' formative assessment capacities derived from undergraduate topics or courses. The same observation might apply to pre-licensure master's degree programs although the distinction between the two forms of master's degrees was not sufficiently clear to warrant the same level of concern based on the results of this study. Without graduate degrees, about half of all teachers may be conducting the necessary processes without the benefit of any memorable or value-added educational instruction in standards, assessments, formative processes, and methods that inform or differentiate instruction. While professional development and training can address these needs and goals, reliance on topic specific training alone for what is needed is insufficient (McMillan, 2003). An effective, education-based introduction to the concepts involved and their systems-like nature are essential. Understanding the niche of each component through a course with the elements examined in this study, supported by real-world standards, lesson plans, assessments, tests, and data sets might enable the development of courses and included exercises that provide the needed *memorable*

experience. It is therefore recommended that consideration be given to the inclusion of required topical materials or a course that presents the issues of and the processes for assessment and formative methods and testing in conjunction with instruction and planning. To these, introduction to the analytic skills required to benefit from the availability of testing data should be added to pre-licensure preparation programs. While such a recommendation challenges the content of existent, tightly-packed programs, the priority of needs amongst all program content might suggest that a reprioritization of included materials is in order.

For in-service teachers, experience in their respective classrooms brings the relevance of the selected topics to life. The examination of testing data, their analyses, and immediate need to amend instruction become real concerns. The preliminary analyses of the professional development component's variables suggest that a level of participation in specific topics is most beneficial as those who had acquired the highest levels of training had students who attained the highest levels of achievement. It is recommended that professional development and training in specific topics be required. If the data and its analyses are correct, students' achievements would eventually rise.

Considering the affirmation of professional development and training noted in the literature and the relevance of professional development and training on teachers' capacities to use formative testing results to inform instruction concluded in this study, it is recommended that greater fusion of the benefits and strengths of professional development with the strengths of education be sought. Each form of instruction has something to contribute. It is unfortunate that the two seem to be separate, uncoordinated efforts endeavoring to contribute rather than be components of a single system dedicated

to the best possible achievement by all students. While the professional development school concept discussed in the literature may not be the answer, it is recommended that partnership solutions that provide a systematic approach be examined between colleges of education and public school districts or divisions.

Finally, based on the district studied, the assignment of teachers with non-educational field undergraduate foci seems to disadvantage some schools. The preliminary analyses suggested that the percentage of non-educational field undergraduates in Group C schools was higher than in either Groups A and B. Whether these assignments have any influence on the achievement noted is unknown. As a matter of policy, however, it would seem that monitoring the mix of more seasoned, elementary and early childhood educated teachers, such as those noted in Group A schools, with teachers whose pre-licensure preparation was in other disciplines, might be a point for consideration when seeking to improve overall student achievement. This topic is worthy of further review and study if balancing of teacher capabilities across a district's schools is of interest.

For Future Investigations

Many lessons have been learned through the conduct of this study. There appear to be other variables, such as socioeconomics or a more complete definition of teacher experiences with regard to graduate education, at work. In seeking further clarity, the following are recommendations for future investigations based on the outcomes and experiences of this study.

Generally, the first recommendation for future investigations includes the improvement of the design and methods used to conduct this study. The linkage of

principal component analysis components scores to student achievement seems to be the proper final outcome and likely follow-on to this study as it might provide insights capable of producing generally higher achievement in all content areas. The finding that Group A teachers, in 13 of 15 variables extracted in three components, were favored, despite a class size disadvantage, suggests that the this study may have value. The ambiguity of the remaining 2 of 15 variables and relative rankings of Groups B and C in all variables warrant further study. Before doing so, however, it seems appropriate to repeat this study with a broader sampling of teachers and an appropriately revised design. For example, rather than excluding teachers based on instructional and testing models of their schools, these attributes should be made variables describing a teacher and including them in the study. More precision in the collecting and reporting of instructional hours and topical experiences would potentially eliminate noted ambiguity. A clearer differentiation in the types of degree programs experienced, in undergraduate, pre-licensure master's degrees, advanced studies master's programs, and other graduate work, seems appropriate. A measure of the years of experience with the content of programs attended and degrees achieved seems an element that might contribute to greater clarity. Inclusion of the dimension provided by increasing specialization in elementary content might contribute to greater clarity as might the effects of testing models other than district only testing, provided that evaluations of the tests used can be made. In professional development and training, clearer and/or more detailed definition of training participated in might clarify differences. The use of potential co-variants, such as mathematics specialists, socioeconomics by various measures, and other school factors requires more study and specification. The second recommendation is to conduct the revised study on a

district-wide basis.

Regarding design orientation, this study looked at teachers associated with schools then ranked the schools in order to stratify their differences. The reasoning was to look at a nested design with school and individual effects. That reasoning is debatable. The alternative design proposed changes the level of analyses to that of looking at the individual teacher with their schools as a variable instead of schools with teachers as a variable. In this situation, effects caused by teachers independent of schools may lead to more conclusive relationships and results. It is possible that re-running analyses with the current data in a new design would not necessarily result in a different outcome but might produce more meaningful components.

The criterion for this study involved teachers teaching four core subjects to a classroom of third grade students. In hindsight, this criterion supports thinking of the past. It is recommended that an expanded study, such as this, be conducted involving all teachers. Teacher-subject selection should be independent of grade of assignment, content area of interest, model of instruction employed, or periodicity of school-wide testing although each of these should be captured as a series of variables. These data should be collected as part of the interview – survey process as should ethnicity and socioeconomic data. The precision of the protocol used for interview should be expanded to allow multiple items for all variables of interest in order to better test for internal reliability. As a case in point, the protocol from which this study's protocol was derived was eight times the length and probed courses and content with more definition while containing the elements necessary for assessing all internal reliability. To be more accurate, those who voluntarily participate in such studies often represent strata unto

themselves when compared to those who choose not to participate. The latter represents an unknown quantity. Such is the case in this study. There were more low achieving schools not part of this study than high performing schools.

Regarding curriculum specialist effects, more research is recommended. There seemed to be a substantial range of agreement or disagreement between teacher and curriculum coordinator evaluations of the specialists. The criterion on which specialist evaluations were based was not specified though expected to be professional. An r of 0.608, considering the range of evaluations provided, seemed to suggest a similarity with some semblance of a standard. However, an r of -0.656 seemed to suggest completely different criteria. There is some interest in the observation that high and low performing teachers seemed to agree while median performing teachers did not. These differences may represent mismatches in expectation or the depth of support received. One might hypothesize that there is an independence from or a dependency on the specialist that warrants investigation, especially since it may have an effect of achievement outcomes.

Finally, the interest of the research was the formative testing to inform instruction process. Such a process is independent of a content area. Any study examining education or training contributions to achievement must also be attentive to the teacher's interest in and preparation to teach all subjects taught.

REFERENCES

Airasian, P. W. (1988). Measurement driven instruction: a closer look. *Educational Measurement: issues and practice, 7*, 6-11.

Almond, R. G., Steinberg, L., & Mislevy, R. J. (2003). *A four-process architecture for assessment delivery, with connections to assessment design* (CSE Report 616). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Anderson, L., & Olsen, B. (2006). Investigating early career urban teachers' perspectives on and experiences in professional development. *Journal of Teacher Education, 57*(4), 359-377.

Anderson, B., & Postl, B. (2001, April). *Using large data sets as a basis for school improvement.* Paper presented at the Annual Meeting of the American Eduational Research Association, Seattle, WA.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing - a revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Ash, D., & Levitt, K. (2003). Working within the zone of proximal development: formative assessment of professional development. *Journal of Science Teacher Education 14*(1), 23-48.

Ashlock, R. B. (1994). *Error patterns in computation* (6th ed.). New York: Macmillan Publishing Company.

Assessment Reform Group. (1999). *Assessment for learning: beyond the black box.* Cambridge, UK: Cambridge University, School of Education.

Assessment Reform Group. (2002). *Testing, motivation and learning.* Cambridge, UK: Cambridge University, School of Education.

Association for Childhood Education International. (2003). *Elementary education standards*. Olney, MD: Author.

Athanases, S. Z., & Achinstein, B. (2003). Focusing new teachers on individual and low performing students: the centrality of formative assessment in the mentor's repertoire of practice. *Teacher College Record, 105*(8), 1486-1520.

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matrices: the case of self-reported affect and perceptions at work. *Journal of Applied Psychology, 75*(5), 547-560.

Baker, E. L. (1994). Researchers and assessment policy development: a cautionary tale. *American Journal of Education, 102*(4), 450-477.

Baker, E. L. (2003). *Multiple measures: toward tiered systems* (CSE Report 607). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Linn, R. L. (1997). *Emerging educational standards of performance in the United States* (CSE Technical Report 437). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Linn, R. L. (1998). *Back to basics: indicators as a system.* CRESSTline. Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for

performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond and G. Sykes (Eds.), *Teaching as the learning profession* (pp. 3–31). San Francisco, CA: Jossey-Bass.

Bandalos, D., & Boehm, M. (2007). *Four common misconceptions in exploratory factor analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Ill.

Bangel, N. J., Capobianco, B., & Moon, S. M. (2006). Professional development of preservice teachers: teaching in the super Saturday Program. *Journal for the Education of the Gifted, 29*(3), 339-361.

Barlow, A. T., & Reddish, J. M. (2006, winter). Mathematical myths: teacher candidates' beliefs and implications for teacher educators. *The Teacher Educator, 41*(3), 145-157.

Barr, R., & Dreeben, R. (1977). Instruction in classrooms. *Review of Research in Education, 5*, 89-162.

Bass, K. M., & Glaser, R. (2004). *Developing assessment to inform teaching and learning* (CSE Technical Report 628). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bedwell, L. E. (2004). *Data-driven instruction*. Bloomington, IN: Phi Delta Kappan Educational Foundation.

Berk, R. A. (Ed). (1980). *Criterion-referenced measurement: the state of the art.*

Baltimore, MD: The John Hopkins University Press.

Black, P. (1998, March). Learning, league tables and national assessment: opportunity lost or hope deferred. *Oxford Review of Education, 24*(1), 57-68.

Black, P. (2000, Sep-Dec). Research and the development of educational assessment. *Oxford Review of Education*, 26(3/4), 407-419.

Black, P., & Wiliam, D. (1998a, March). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Black, P., & Wiliam, D. (1998b, October). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-144, 146-148.

Bliem, C. L., & Davinroy, K. H. (1997). *Teachers' beliefs about assessment and instruction in literary* (CSE Technical Report 421). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Block, J. H., & Burns, R. B. (1976). Mastery learning. *Review of Research in Education, 4*, 3-49.

Bloom, B. S. (Ed)., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives, handbook I: cognitive domain*. New York: David McKay Company, Inc.

Borko, H. (Autumn, 1997). New forms of classroom assessment: implications for staff development. *Theory into Practice, 36*, 231-238.

Borko, H. (2004). Professional development and teacher learning: mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Borko, H., Mayfield, V., Marion, S. F., Flexer, R. J., & Cumbo, K. (1997). *Teachers'*

developing ideas and practices about mathematics performance assessment: successes, stumbling blocks, and implications in professional development (CSE Technical Report 423). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Borko, H., Stecher, B. M., Alonzo, A., Moncure, S., & McClam, S. (2003). *Artifact packages for measuring instructional practice: a pilot study* (CSE Report 615). Los Angeles, CA: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bracey, G. W. (1987). Measurement-driven instruction: catchy phrase, dangerous practice. *Phi Delta Kappan, 69*(9), 683-686.

Briars, D. J., & Resnick, L. B. (2000). *Standards assessments - and what else? the essential elements of standards-based school improvement* (CSE Technical Report 528). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Brimijoin, K., Marquissee, E., & Tomlinson, C. A. (February, 2003). Using data to differentiate instruction. *Educational Leadership, 60*(5), 70-73.

Brown, K., & Capp, R. (2003). Better data for better learning. *Leadership, 33*(2), 18-39.

Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2002). District accountability without a state assessment: a proposed model. *Educational Measurement: Issues and Practice, 21*(4), 6-16.

Burnaford, G., & Hobson, D. (1995). Beginning with the group: collaboration as the cornerstone of graduate teacher education. *Action in Teacher Education, 17*, 67-

75.

Cady, J., Meier, S. L., & Lubinski, C. A. (2006). Developing mathematics teachers : the transition from preservice to experienced teacher. *The Journal of Educational Research, 99*(5), 295-205.

Catwali, G. (2003). Lessons from research that changed education. *Educational Leadership, 60*(5), 18-21.

Center for Research on Education, Diversity, and Excellence. (2002). Glossary. Retrieved November 16, 2006, from http://crede.berkeley.edu/tools /glossary.html

Childs, R. A., Ross, M., & Jaciw, A. P. (2002). Initial teacher certification testing: preservice teachers' experiences and perceptions. *Canadian Journal of Education, 27*(4), 455-475.

Cizek, G. J. (1993). Rethinking psychometricians' beliefs about learning. *Educational Researcher, 22*(4), 4-9.

Cizek, G. J. (1998). The assessment revolution's unfinished business. *Kappa Delta Pi Record, 34*(4), 144-149.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. Educational *Measurement: Issues and Practice, 19*(2), 16-23.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004, Winter). Setting performance standards: contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Clark, C., & Peterson, P. (1986). Teachers' thought processes. In M. Wittrock (Ed.),

Handbook of research on teaching (3rd ed.). New York: Macmillan.

Clark, S. L., & Weismantle, M. (2003). *Employment status: 2000* (C2KBR-18). Washington, DC: U. S. Census Bureau.

Cliff, N., & Hamburger, C. D. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin, 68*, 430-445.

Cochran-Smith, M. (2003). Assessing assessment in teacher education. *Journal of Teacher Education, 54*(3), 187-191.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report Series RR-43). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Cooney, T. J., & Wiegel, H. G. (2003). Examining the mathematics in mathematics teacher education. In A. Bishop, M. Clements, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Second handbook of mathematics education* (pp. 795-828). Boston, MA: Kluwer Academic.

Costa, A. L. (February, 1993). How world-class standards will change us. *Educational Leadership, 50*(5), 50-51.

Council of the Great City Schools. (2003). *Charting a new course for the Richmond public schools*. Washington, DC: Council of the Great City Schools.

Council of the Great City Schools. (2004a). *Restoring Excellence to the District of Columbia Public Schools*. Washington, DC: Council of the Great City Schools.

Council of the Great City Schools. (2004b). *Beating the odds*. Washington, DC: Council of the Great City Schools.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.*

Belmont, CA: Wadsworth Group.

Crocker, L. (2003). Teaching for the test: validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*(3), 5-11.

Cromey, A. (2000). Using student assessment data: what can we learn from schools? Policy Issues, Issue 6. (ERIC Document Reproduction Service No. ED452593).

Cunningham, G. K. (1986). *Educational and psychological measurement.* New York: Macmillan Publishing Company.

Dahll f, U. (1971). *Ability grouping, content validity and curriculum process analysis.* New York: Teachers College Press.

Darling-Hammond, L. (1994a, Spring). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5-30.

Darling-Hammond, L. (1994b, August). National standards and assessment: will they improve education? *American Journal of Education, 102*(4), 478-510.

Darling-Hammond, L. (2004, June). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047-1085.

Darling-Hammond, L., & Ball, D. L. (1998). *Teaching for high standards: what policymakers need to know and be able to do.* National Commission on Teaching & America's Future. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world what teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

Delandshere, G. (1996, November). From static and prescribed to dynamic and principled

assessment of teaching. *The Elementary School Journal, 97*(2), 105-120.

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record, 104*(7), 1461-1484.

Delandshere, G., & Jones, J. (1999, Spring). Elementary teachers' beliefs about assessment in mathematics: a case of assessment paralysis. *Journal of Curriculum and Supervision, 14*(3), 216-240.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14-24.

DeMary, J. (personal communication, July 22, 2005a) [Supts. Memo No. 142]. Transmittal of the *2005 standards of quality*. Retrieved November 12, 2005 from http://www.doe.virginia.gov /VDOE /suptsmemos/2005.

DeMary, J. (personal communication, August 5, 2005b) [Supts. Memo No. 149]. *Career and technical education federal Perkins act funds for July 1, 2005, through June 30, 2006*. Retrieved November 12, 2005 from http://www.doe.virginia.gov /VDOE /suptsmemos/2005.

DeNavas-Walt, C., Proctor, B. D., & Mills, R. J. (2004). *Income, poverty, and health insurance coverage in the United States: 2003* (P60-226). Washington, DC: U. S. Census Bureau.

Doheny, C. (2002). Graduate teacher education and change: insights from a first-grade teacher's experience. *Action in Teacher Education, 24*(3), 43-50.

Ebel, R. L. (1968). The value of internal consistency in classroom examinations. *Journal of Educational Measurement, 5*(1), 71-73.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. *Journal of Educational Measurement, 19*(4), 267-278.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Eisner, C. (Ed.). (2000). *A report on early lessons learned in the efforts to end social promotion in the nation's public schools.* Washington, DC: U. S. Department of Education and Council of the Great City Schools.

Eisenberg, A. F. (1999). Forms of socialization: graduate education and the graduate seminar. Response to "rethinking the graduate seminar." *Teaching Sociology, 27*(2), 187-191.

Engelhart, M. D. (1950). Examinations. In W. S. Monroe (Ed.), *Encyclopedia of educational research* (pp. 407-414). New York: Macmillan.

Esposito, J., & Smith, S. (2006). From reluctant teacher to empowered teacher-researcher one educator's journey toward action research. *Teacher Education Quarterly, 33*(3), 45-60.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Thousand Oaks: Sage Publications, Inc.

Frederiksen, N. (1994). The integration of testing with teaching: applications of cognitive psychology in instruction. *American Journal of Education, 102*(4), 527-564.

Gerges, G. (Fall, 2001). Factors influencing preservice teachers "variation in use of instructional methods: why is teacher efficacy not a significant contributor?" *Teacher Education Quarterly, 28*(4), 71-88.

Gilbert, S. L., & Smith, L. C. (2003). A bumpy road to action research. *Kappa Delta Pi*

Record, 39(2), 80-83.

Glaser, R. L. (1962). Psychology and instructional technology . In R. L. GLASER (Ed.), *Training research and education* (pp. 559-78) . Pittsburgh, PA: University of Pittsburgh Press.

Glaser, R., & Silver, E. (1994). *Assessment, testing, and instruction: retrospect and prospect* (CSE Technical Report 379). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Good, T. L., & Brophy, J. E. (2000). *Looking in classrooms* (8th ed.). New York: Longman.

Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Boston, MA: Pearson Education.

Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed.). Boston, MA: Pearson Education.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan Publishing Company.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265-275.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*(4), 347-354.

Haertel, E. H. (1999, May). Performance assessment and education reform. *Phi Delta Kappan, 80*(9), 662-666.

Haertel, E. H., & Herman, J. (2005, June). *A historical perspective on validity arguments for accountability testing* (CSE Report 654). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement scores and the origins of test score pollution. *Educational Researcher 20*(5), 2-7.

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., & Bugliari, D. (2003, Spring). Studying large-scale reforms of institutional practice: an example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*(1), 1-29.

Heiss, A. M. (1968). Graduate education today: an instrument for change?: the answer seems to be no. *The Journal of Higher Education, 39*(1), 1-10.

Heathcott, J. (2005, November-December). Trained for nothing. *Academe.* Retrieved October 17, 2006, from http://www.aaup.org/publications/academe/2005/nd /feat/heat.htm

Heritage, M., Lee, J., Chen, E., & LaTorre, D. (2005, September). *Upgrading America's use of information to improve student performance* (CSE Report 661). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. (1997). Assessing new assessments: how do they measure up? *Theory into Practice, 36*, 196-204.

Herman, J. L., Brown, R. S., & Baker, E. L. (2000). *Student assessment and student achievement in the California public school system* (CSE Technical Report 519). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L., & Winters, L. (1994). Portfolio research: a slim collection. *Educational Leadership, 52*(2), 48-55.

Hiebert, J. (1999). Relationships between research and the NCTM standards. *Journal for Research in Mathematics Education, 30*(1), 3-19.

Hughes, G. B. (1999). Facilitating the development of preservice teachers in a climate of reform: lessons learned from mathematics and assessment reform. *The Journal of Negro Education 68*(3), 352-365.

Impara, J. C., Plake, B. S., & Fager, J. J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into Practice, 32*(2), 113-117.

Interstate New Teachers Assessment and Support Consortium (INTASC). (1992). *Model Standards for Beginning Teacher Licensing and Development: A Resource for State Dialogue.* Washington, DC: Council of Chief State School Officers.

Johnson, M., & Button, K. (2000). Connecting graduate education in language arts with teaching contexts: the power of action research. *English Education, 32*(2), 107-126.

Johnston, D. A., & Lawrence, J. T. (2004). Using data to inform instruction. *Leadership, 34*(2), 28-29, 35.

Johnston, P., Guice, S., Baker, K., Malone, J., & Michelson, N. (1995). Assessment of teaching and learning in "literature-based" classrooms. *Teaching & Teacher*

Education, 11(4), 359-371.

Jones, M. G., & Vesilind, E. (1995). Preservice teachers' cognitive frameworks for class management. *Teaching and Teacher Education, 11*(4), 313-330.

Kachigan, S. K. (1986). *Statistical analysis an interdisciplinary introduction to univariate & multivariate method.* New York: Radius Press.

Kelley, C., Heneman, H., & Milanowski, A. (2000). *School-based performance award programs, teacher motivation, and school performance: findings from a study of three programs* (CPRE Research Report Series RR-44). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Khanna, R., Trousdale, D., Penuel, W. R., & Kell, J. (1999, April). *Supporting data use among administrators: results from a data planning model.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Kilpatrick, J. (1992). A history of research in mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 3-38). New York: Macmillan.

Kim, M. M., Andrews, R. L., & Carr, D. L. (2004). Traditional versus integrated preservice teacher education curriculum a case study. *Journal of Teacher Education, 55*(4), 341-356.

Kittock-Sargent, M. (1995, April 12). Standards: setting the standards from state to state. *Education Week on the WEB.* Retrieved March 12, 2005, from http://www.edweek.org/ew/articles/1995/04/12/29stcht.h14.html

Klein, D. C. D., O'Neil, H. F., Jr, & Baker, E. L. (1998). *A cognitive demands analysis of*

innovative technologies (CSE Technical Report 454). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37*(4), 752-777.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND. (MR- 1014-EDU).

Koretz, D. M., & Berends, M. (2001). *Changes in High School Grading Standards in Mathematics, 1982-1992*. (RAND MR-1445-CB). Santa Monica, CA: RAND.

Koretz, D. M., McCaffery, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., McCaffery, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont portfolio assessment program* (CSE Technical Report 355). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: perceived effects of the Maryland school assessment program* (CSE Technical Report 409). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D., Stecher, B., Klein,S., McCaffery, D., & Deibert, E. (1993). *Can portfolios*

assess student performance and influence instruction? The 1991-92 Vermont Experience (CSE Technical Report 371). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives handbook II: affective domain.* New York: David McKay Company, Inc.

Lewis, A. (2001, October). *Educational Accountability in the 21st Century* (CSE Technical Report 549). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lewis, A. (2005, July). *Research guidance: assessment, accountability, action!* (CSE Technical Report 658). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L. (1994a, August). Evaluating the technical quality of proposed national examination systems. *American Journal of Education, 102*(4), 565-580.

Linn, R. L. (1994b, December).Performance assessment: policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Linn, R. L. (1998). *Assessment and accountability* (CSE Technical Report 490). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L. (2000, March). Assessment and accountability. *Educational Researcher,*

29(2), 4-16.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Technical Report 539). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Linn, R. L., Baker, E. L., & Betebenner, D. (2002). *Accountability systems: implications of requirements of the no child left behind act of 2001* (CSE Technical Report 567). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L., & Haug, C. (2002). *Stability of school building accountability scores and gains* (CSE Technical Report 561). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lissitz, R. W. (2005, November). *Director's opening remarks at the meeting of MARCES conference on longitudinal modeling of student achievement,* College Park, MD.

MacDonald-Ross, G. (1993). Socrates versus Plato: the origins and development of Socratic thinking. *Aspects of Education 49*, 9-22.

Madaus, G. F. (1985). Public policy and the testing profession: you've never had it so good? *Educational Measurement: Issues and Practice, 4*(4), 5-11.

Madaus, G. F. (Spring, 1988). The distortion of teaching and testing: high-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Madaus, G. F., Kellaghan, T., Rakow, E. A., & King, D. (1979). The sensitivity of measures of school effectiveness. *Harvard Educational Review, 49*(2), 207-230.

Madaus, G. F., & O'Dwyer, L. M. (1999, May) A short history of performance assessment. *Phi Delta Kappan, 80*(9), 688-695.

Marzano, R. J. (1992). *A different kind of classroom: teaching with dimensions of learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. (2000). Implementing standards-based education. *Teacher Librarian, 28*(2), 30-32.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press.

Marzano, R. J. (2003). Using data: two wrongs and a right. *Educational leadership, 60*(5), 56-60.

Massell, D. (1998). *State strategies for building capacity in education: progress and continuing challenges* (CPRE Research Report Series RR-41). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

McMillan, J. H. (2001, Spring). Secondary teachers' classroom assessment and grading practices. *Educational Measurement, Issues and Practice 20*(1), 20-32.

McMillan, J. H. (2003, Winter). Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34-43.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teacher's classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203-213.

McNair, S., Bhargava, A., Adams, L., Edgerton, S., & Kypros, B. (2003). Teachers speak

out on assessment practices. *Early Childhood Education Journal, 31*(1), 23-31.

McTighe, J., & Thomas, R. S. (2003). Backward design for forward action. *Educational Leadership, 60*(5), 52-55.

Meyers, L. S., Gamst, G., & Guario, A. J. (2006). *Applied multivariate research.* Thousand Oaks: Sage Publications.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Measurement, 33*(4), 379-416.

Mislevy, R. J. (2003). *Argument substance and argument structure in educational assessment* (CSE Report 605). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Breyer, F. J., & Johnson, L. (2001). *Making sense of data from complex assessments* (CSE Technical Report 538). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Steinberg, L. S., Almond, R., G., Haertel, G. D., & Penuel, W. R. (2001). *Leverage points for improving educational assessment* (CSE Technical Report 534). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2002). *Psychometric principles in student assessment* (CSE Technical Report 583). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Moss, P. A. (Autumn, 1992). Shifting conceptions of validity in educational

measurement: implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A. (January/February, 1996). Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-28, 43.

Myers, L. S., Gamst, G., & Guarion, A. J. (2006). *Applied multivariate research design and interpretation.* Thousand Oaks, CA: Sage Publications.

National Association for the Education of Young Children. (2001). *Standards for early childhood professional preparation initial licensure programs.* Washington, D. C.: Author.

National Association for the Education of Young Children. (2002). *Standards for early childhood professional preparation advanced programs.* Washington, D. C.: Author.

National Council for the Accreditation of Teacher Education. (1998). *Program for initial preparation of K-4 teachers with an emphasis in mathematics, 5-8 mathematics teachers, 7-12 mathematics teachers.* Reston, VA: National Council of Teachers of Mathematics.

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

NCS Pearson. (2005). *Prosper assessment system.* [Computer software]. Available at http://www.pearsonncs.com/prosper/index.htm.

Neill, D. M. (1997, September). Transforming student assessment. *Phi Delta Kappan, 79*(1), 34-40, 58.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7.

Noble, A. J., & Smith, M. L. (1994). Old and new beliefs about measurement-driven reform: "build it and they will come". *Educational Policy, 8*(2), 111-136.

Norfolk Public Schools. (2005). *School board performance report 2003-2004*. Norfolk, VA.

Norfolk Public School. (2004). *Performance report - elementary school - 008*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 009*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 015*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 018*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 019*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 021*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 022*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 023*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 024*. Retrieved

April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 026.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 027.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 028.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 029.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 031.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 034.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 035.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 036.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 037.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 041.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 042.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 045.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 046.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 050.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 051.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 052.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 053.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 054.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 056.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 059.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 060.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 061.* Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 065.* Retrieved

April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 068*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 069*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 106*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Norfolk Public School. (2004). *Performance report - elementary school - 114*. Retrieved April 8, 2005 from http://cf.nps.k12.va.us/performes/.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

O'Neil, J. (1993). On the new standards project: a conversation with Lauren Resnick and Warren Simmons. *Educational Leadership, 50*(5), 17-21.

Oregon Public Education Network. (2004). *Glossary*. Retrieved from Retrieved February 8, 2007 from http://open.k12.or.us/glossary.html

O'Sullivan, R. G., & Chalnick, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17-19.

Otero, V. K. (2006). Moving beyond the "get it or don't" conception of formative assessment. *Journal of Teacher Education, 57*(3), 247-255.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Pelligrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: linking theories of cognition and learning with assessment and

instructional practice. *Review of Research in Education, 24*, 307-353.

Popham, W. J. (Ed.). (1971). *Criterion-referenced measurement an introduction.* Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (1987a). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*(9), 679-682.

Popham, W. J. (1987b). Muddle-minded emotionalism. *Phi Delta Kappan, 68*(9), 687-688.

Popham, W. J. (1998, January). Farewell, curriculum: confessions of an assessment convert. *Phi Delta Kappan, 79*(5), 380-384.

Popham, W. J. (1999). Where large scale assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*(3), 13-17.

Popham, W. J. (1999). *Modern educational measurement: practical guidelines for educational leaders* (3rd ed.). Boston, MA: Allyn & Bacon, Inc.

Popham, W. J. (2004). *Classroom assessment: what teachers need to know.* (4th ed.). Boston, MA: Allyn & Bacon.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (May, 1985). Measurement-driven instruction: it's on the road. *Phi Delta Kappan, 66*(9), 628-634.

Prawat, R. S. (1992). Teachers' beliefs about teaching and learning: a constructivist perspective. *American Journal of Education, 100*(3), 354-395.

Putnam, R., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *The international*

handbook of teachers and teaching (pp. 1223–1296). Dordrecht, The Netherlands: Kluwer.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.

Ravitch, D. (1996). The case for national standards and assessments. *The Clearing House, 69*, 134-135.

Ravitz, J. (2002). CILT2000: Using Technology to Support Ongoing Formative Assessment in the Classroom. *Journal of Science Education and Technology, 11*(3), 293-296.

Reeves, D. B. (2002). *The leader's guide to standards a blueprint for educational equity and excellence.* San Francisco, CA: Jossey-Bass.

Reigeluth, C. M. (1997). Educational standards to standardize or to customize learning? *Phi Delta Kappan, 79*, 202-206.

Reis, S. M., & Westberg, K. L. (1994). The impact of staff development on teachers' ability to modify curriculum for gifted and talented students. *Gifted Child Quarterly,* 38, 127-135.

Resnick, L. B. (1994). Performance puzzles. *American Journal of Education, 102*(4), 511-526.

Resnick, L. B., & Harwell, M. (2000). *Instructional variation and student achievement in a standards-based education district* (CSE Technical Report 522). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Resnick, L. B., Nolan, K. J., & Resnick, D. P. (1995). Benchmarking education standards. *Educational Evaluation and Policy Analysis, 17*(4), 438-461.

Richardson, V. (2003). Preservice teacher beliefs. In J. Raths & A. C. McAninch (Eds.), T*eacher beliefs and classroom performance: the impact of teacher education.* (pp. 1-22). Greenwich, CT: Information Age Publishing.

Rigney, S. L., & Martineau, J. A. (2005, November). *NCLB and growth models: in conflict or concert?* Paper presented at the meeting of MARCES Conference on Longitudinal Modeling of Student Achievement, College Park, MD.

Rogers, E. M. (1995). *Diffusion of innovation* (4th ed.). New York: The Free Press.

Rubenstein, M.C., & Wodatch, J. K. (2000). *Stepping up to the challenge: case studies of educational improvement in title I secondary schools.* U.S. Department of Education, Office of the Under Secretary, Planning, and Evaluation Service, Elementary and Secondary Education Division, Washington, DC.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston: Northwestern University Press.

Sachs, S. K. (March/April, 2004). Evaluation of teacher attributes as predictors of success in urban schools. *Journal of Teacher Education, 55*(2), 177-187.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Sax, C., & Fisher, D. (2001). Using qualitative action research to effect change: implications for professional education. *Teacher Education Quarterly, 28*(2), 71-80.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3-6, 12.

Schafer, W. D. (1993, Spring). Assessment literacy for teachers. *Theory into Practice, 32*(2), 118-126.

Schmoker, M., & Marzano, R. J. (1999). Realizing the promise of standards-based education. *Educational Leadership, 56*(6), 17-21.

Shepard, L. A. (1980). Technical issues in minimum competency testing. *Review of Research in Education, 8*, 30-82.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2-16.

Shepard, L. A. (2000a). *The role of classroom assessment in teaching and learning* (CSE Technical Report 517). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed). Washington, DC: American Educational Research Association.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Baratz-Snowden, J., Gordon, E., Gutierrez, C., & Pacheo, A. (2005). Assessment. In L. Darling-Hammond & Bransford, J. (Eds.), *Preparing Teachers for a Changing World what teachers should learn and be able to do*. San Francisco: Jossey-Bass.

Shulman, L. S., & Elstein, A. S. (1975). Studies of problem solving, judgment, and decision making: implications for educational research. *Review of Research in Education, 3*, 2-42.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50*(5), 11-15.

Skinner, B. F. (1963). Operant behavior. *American Psychologist, 18*(8), 503-515.

Sleight, D. A. (1993). *A developmental history of training in the United States and Europe*. Retrieved March 12, 2005 from Michigan State University web site: http://www.msu.edu/~slightd/trainhst.html.

Sloane, F. C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory into Practice, 42*(1), 12-17.

Smith, M. L., Noble, A. J., Heinecke, W., Seck, M., Parish, C., Cabay, M., Junker, S. C., Haag, S., Tayler, K., Safran, Y., Penley, Y., & Bradshaw, A. (1997). *Reforming schools by reforming assessment: consequences of the Arizona student assessment program (ASAP): equity and teacher capacity building* (CSE Technical Report 425). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Smith, M. L., & Shepard, L. A. (1988). Kindergarten readiness and retention: a qualitative study of teachers' beliefs and practices. *American Educational Research Journal, 25*(3), 307-333.

Smith, R. L., & Yen, W. M. (2005, November). *Models for evaluating grade-to-grade growth*. Paper presented at the meeting of MARCES Conference on Longitudinal Modeling of Student Achievement, College Park, MD.

Smith, T. M., & Rowley, K. J. (2005). Enhancing commitment or tightening control: the function of teacher professional development in an era of accountability. *Educational Policy, 19*(1), 126-154.

Software America. (2005). *Assessor*. [Internet software]. Retrieved March 15, 2005 from http://www.softwareamerica.com/assessor.

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and

perceptions at work: myth or significant problem? *Journal of Applied Psychology, 72*(3), 438-443.

Spector, P. E. (1992). *Summated rating scale construction an introduction.* Thousand Oaks: Sage Publications, Inc.

Spicer, J. (2005). *Making sense of multivariate data analysis.* Thousand Oaks: Sage Publications, Inc.

Spielmann, G., & Radnofsky, M. L. (2001). Learning language under tension: new directions from a qualitative study. *The Modern Language Journal, 85*(2), 259-278.

Spillane, J. P. (2000). *District leaders' perceptions of teacher learning* (CPRE Occasional Paper Series OP-05). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Stanford, R. L. (2001). Some design considerations for masters degree curricula in elementary and secondary education. *Action in Teacher Education, 23*(2), 1-8.

Stecher, B. M., Barron, S., Borko, H., & Wolf, S. (1997). *Important features of state assessment systems from the local perspective: interim report* (CSE Technical Report 472). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (CSE Technical Report 525). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-*

based assessment on classroom practices: results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing (CSE Technical Report 482). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12.

Stiggins, R. J. (2001a). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall

Stiggins, R. J. (2001b). The unfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*(3), 5-15.

Stiggins, R. J. (2002). Assessment crisis: the absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 86*(1), 22-27.

Stiggins, R. (2005). From formative assessment to assessment for learning: a path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324-328.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.

Strauss, R. (August 25, 2003). *District quarterly assessments - purpose*. Norfolk, VA: Norfolk Public Schools.

Stotsky, S. (2006). Who should be accountable for what beginning teachers need to

know? *Journal of Teacher Education, 57*(3), 256-268.

Sutton, R. E., Cafarelli, A., Lund, R., Schurdell, D., & Bichsel, S. (1996). A developmental constructivist approach to pre-service teachers' ways of knowing. *Teaching & Teacher Education, 12*(4), 413-427.

Swan, M. (2006). Designing and using research instruments to describe the beliefs and practices of mathematics teachers. *Research in Education, 75*, 58-70.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

Taylor, C. (1994, Summer). Assessment for measurement or standards: the peril and promise of large-scale assessment reform. *American Educational Research Journal, 31*(2), 231-262.

Thorndike, R. M. (1997). *Measurement and evaluation is psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

United States Code. (2004). Title IX, 20 U.S.C. 7801.

United States Congress. (1994). *The improving America's schools act of 1994*. (Public Law 103-382). Washington, DC: Government Printing Office.

United States Congress. (2002). *No child left behind act of 2001* (Public Law 107-110). Washington, DC: Government Printing Office.

United States Census Bureau. (2000). *Profile of general demographic characteristics: Norfolk City, Virginia*. DP-1. Retrieved April 17, 2005 from http://factfinder.census.gov.

United States Census Bureau. (2002). *Census 2000 demographic profile highlights: Norfolk City, Virginia*. DP-1. Retrieved June 11, 2005 from http://factfinder.census.gov.

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies, 34*(2), 201-225.

Virginia Board of Education. (April 25, 1996). *Virginia board of education resolution establishing policies regarding the Virginia state assessment program (resolution 1996-4)*. Richmond, VA: Virginia Board of Education.

Virginia Department of Education. (November 20, 1998). Standards of Learning (SOL) test scaled scores and equating. Superintendent's Memo 179. Richmond, VA: Virginia Department of Education.

Virginia Department of Education. (2002). *SY 2002-2003 Free and reduced lunch program eligibility repor*t. Richmond, VA: Virginia Department of Education.

Virginia Department of Education. (2004). *SY 2004-2005 Free and reduced lunch program eligibility report*. Richmond, VA: Virginia Department of Education.

Virginia Department of Education. (2004). 2005-2006 estimated Perkins Lea allocations grant allocations. Superintendent's Memo 149, attachment A.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge MA: Massachusetts Institute of Technology.

Ward, J. G. (1980). Teachers and testing: A survey of knowledge and attitudes. In L. M. Rudner (ed.). *Testing in our schools*, (pp. 15-24). Washington, D.C.: National Institute of Education.

Ward, M. (1998, April). *A systems approach to middle school evaluation: Guilford county schools' formative approach*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Washington Kids Count. (2001). *Exploring disparities in education achievement: the impact of school funding* (Washington Kids Count Technical Report). Seattle, WA: University of Washington, Human Services Policy Center.

White, C. S., Sturtevant, E. G., & Dunlap, K. L. (2003). Preservice and beginning teachers' perceptions of the influence of high stakes tests on their literacy-related instructional beliefs and decisions. *Reading Research and Instruction, 42*(2), 39-62.

Wiggins, G. (October, 1994). Toward better report cards. *Educational Leadership, 52*(2), 28-37.

Wilen, W., Ishler, M., Hutchinson, J., & Kindsvatter, R. (2004). *Dynamics of effective teaching* (5th ed.). New York: Addison Wesley Longman.

Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.

Wolf, S., Borko, H., McIver, M., & Elliott, R. (1999). *"No excuses": school reform efforts in exemplary schools of Kentucky* (CSE Technical Report 514). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wong, S. P., & McGraw, K. O. (1999). Confidence intervals and f tests for intraclass correlations based on three-way random effects models. *Educational and*

Psychological Measurement, 59(2), 270-288.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: new standards examinations for the California mathematics renaissance* (CSE Technical Report 484). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Yorke, M. (2003). Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*, 477-501.

Zemelman, S., Daniels, H., & Hyde, A. (1998). *Best practice: new standards for teaching and learning in America's schools* (2nd ed.). Portsmouth, NH: Heinemann.

APPENDIX A

Teacher's Questionnaire / Interview Protocol

Random Identification _____

The following questions pertain to your undergraduate preparation program.							
What was your bachelor's degree program major? _____							
If you had a minor, what was (were) the area(s) of focus? _____							
Considering your personal likes and dislikes, where does mathematics rank with history, reading, science, and writing? 1 is most favorite, 5 is least favorite.						1 2 3 4 5	
Did you have mathematics courses beyond the requirements of your program?						Yes	No
If so, what were they? _____ _____ _____							
Did you have any course(s) in which assessment was a topic of discussion?						Yes	No
If so, do you recall how much class time was spent on it?						__ mins hrs crs	
Did you have any course(s) in which testing was a topic of discussion?						Yes	No
If so, do you recall how much class time was spent on it?						__ mins hrs crs	
Did you have any course(s) in which item writing was a topic of discussion?						Yes	No
If so, do you recall how much class time was spent on it?						__ mins hrs crs	
Did you have any course(s) in which data-driven instruction was a topic of discussion?						Yes	No
If so, do you recall how much class time was spent on it?						__ mins hrs crs	
The following questions pertain to your teaching experience.							
How many years have you been teaching in the third grade? _____							
How many years have you been teaching overall? _____							
The following questions pertain to any graduate level education you may have had?							
Have you taken any graduate classes?						Yes	No
What is the major of the degree you seek or have achieved? _____							
Were / are there additional focus areas in your program?						Yes	No
If so, what were / are they? _____ _____ _____							
Have you had graduate courses that contained the following?							
Evaluation	Yes	No	__ hrs	Research design	Yes	No	__ hrs
Statistics	Yes	No	__ hrs	Assessment	Yes	No	__ hrs
Testing	Yes	No	__ hrs	Standards	Yes	No	__ hrs
Test item construction	Yes	No	__ hrs	Analysis of results	Yes	No	__ hrs
Differentiated instruction	Yes	No	__ hrs	Advanced curriculum design	Yes	No	__ hrs
Additional mathematics content	Yes	No	__ hrs	_____	_____	_____	
If you completed your master's degree, when was it awarded? _____							
Do you have graduate level courses beyond your master's degree?							
If so, please indicate the topics included using the format above:							
_____	Yes	No	__ hrs	_____	Yes	No	__ hrs
_____	Yes	No	__ hrs	_____	Yes	No	__ hrs

Random Identification _____

The following questions pertain to professional development, in-service, or capacity development training you may have had. Have you had training in:				
Math methods?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Assessment?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Testing?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Test results analysis?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Data-driven decision-making		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Action research?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Evaluation?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Differentiated instruction?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Strategies, activities, and exercises?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Instructional planning and revision?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
Do you conduct professional development, in-service, or capacity development training as an instructor?		Yes	No	If so, how many total hours?
0 hours	2 or less hours	2 to 5 hours	5 to 10 hours	more than 10 hours
On a scale of 1 to 5, 1 being strong belief and 5 being weak belief, do you believe in formative testing as laid out by the district's 9 week content assessment program?				1 2 3 4 5
On a scale of 1 to 5, 1 being the greatest and 5 being the least, how much assistance do you receive from the math specialist in the analysis of testing results on which subsequent instruction is based?				1 2 3 4 5

Appendix B

Schools' Instructional and Testing Models

School	Random School ID	Instructional Model	Testing Model
School 1			
School 2			
School 3			
School 4			
School 5			
School 6			
School 7			
School 8			
School 9			
School 10			
School 11			
School 12			
School 13			
School 14			
School 15			
School 16			
School 17			
School 18			
School 19			
School 20			
School 21			
School 22			
School 23			
School 24			
School 25			
School 26			
School 27			
School 28			
School 29			
School 30			
School 31			
School 32			
School 33			
School 34			
School 35			

APPENDIX C

Mathematics Curriculum Coordinator's Assessment of Participant Teachers

Teacher	**Random Teacher ID**	**Likert-scaled evaluation of Teachers Capacities (1 – greatest capacity; 5 – least capacity)**
Teacher 1		
Teacher 2		
Teacher 3		
Teacher 4		
Teacher 5		
Teacher 6		
Teacher 7		
Teacher 8		
Teacher 9		
Teacher 10		
Teacher 11		
Teacher 12		
Teacher 13		
Teacher 14		
Teacher 15		
Teacher 16		
Teacher 17		
Teacher 18		
Teacher 19		
Teacher 20		
Teacher 21		
Teacher 22		
Teacher 23		
Teacher 24		
Teacher 25		
Teacher 26		
Teacher 27		
Teacher 28		
Teacher 29		
Teacher 30		
Teacher 31		
Teacher 32		
Teacher 33		
Teacher 34		
Teacher 35		
Teacher 36		
Teacher 37		
Teacher 38		
Teacher 39		

Teacher 40		
Teacher 41		
Teacher 42		
Teacher 43		
Teacher 44		
Teacher 45		
Teacher 46		

APPENDIX D

Mathematics Curriculum Coordinator's Assessment of School Mathematics Specialists

School	Random School ID	Likert-scaled evaluation of Mathematics Specialist (1 – greatest assistance; 5 – least assistance)
School 1		
School 2		
School 3		
School 4		
School 5		
School 6		
School 7		
School 8		
School 9		
School 10		
School 11		
School 12		
School 13		
School 14		
School 15		
School 16		
School 17		

VITA

William Clark Reed

Educational Background

Bachelor of Science, Naval Science; United States Naval Academy, 1969
Master of Science in Education; Old Dominion University, 1996
Doctor of Philosophy, Urban Services, Old Dominion University, 2007

Professional Experience

Assessment Coordinator, Norfolk Public Schools, 2007 to present
Assessment Intern, Norfolk Public Schools, 2003 – 2007
Instructor, Educational Curriculum and Instruction – Instructional Technology, Darden College of Education, Old Dominion University – 2000 – 2003
Graduate Teaching Assistant, Occupational and Technical Studies, Darden College of Education, Old Dominion University – 1996 – 2000
Outstanding Graduate Teaching Assistant, College of Education, 1998
Phi Kappa Phi Honor Society, 1998
Naval Officer, United States Navy, 1969 – 1994
Designated Surface Warfare Officer (1110), 1982; Qualified for Command, 1987
Designated subspecialist in: Command and Control, 1985; Operational Logistics, 1989; and Education and Training Management, 1992
Designated Cryptologic Officer (1610), 1978
Designated subspecialist in Naval Technical Intelligence, 1980
Designated Submarine Officer (1120), 1972
Designated subspecialist in Nuclear Propulsion, 1970

Professional Publications

Reed, W. C., & Ritz, J. M. (2000, April). Design for the future: strategic planning for technology educators [monograph]. [On-line]. International Technology Education Association.

Ritz, J. M., & Reed, W. C. (1999, December). Gaining support for technology education using the art of positioning [monograph]. [On-line]. International Technology Education Association.

Reed, W. C. (1991). Commander, Second Fleet Logistics Operations Tactical Memorandum.

Reed, W. C. (1984). Commander, Second Fleet Airborne Communications Relay (SATCAT) Tactical Memorandum.

Department of Occupational and Technical Studies
Darden College of Education, Room 228
Old Dominion University
Norfolk, VA 23529