AN INVESTIGATION OF TEACHER EDUCATIONAL MEASUREMENT LITERACY

By

CHAD MARTIN GOTCH

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Educational Leadership and Counseling Psychology

AUGUST 2012

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of CHAD MARTIN GOTCH find it satisfactory and recommend that it be accepted.

_____

Brian F. French, Ph.D., Chair

_____

Tariq T. Akmal, Ph.D.

_____

Michael S. Trevisan, Ph.D.

ACKNOWLEDGMENT

This dissertation is the product of the efforts and support of many people. First there are the teachers who gave up their time—and occasionally their pride—to be a part of my studies. I thank my family for unwavering support and interest in what I do, even as understanding what exactly I do has become more and more difficult. I thank those who have served as chairs of my graduate committees—Brian French, Troy Hall, and Cynthia Pierce. You performed tremendously at setting high standards while providing the mentorship and confidence I needed to succeed. Additionally, I thank those individuals who served on my thesis and dissertation committees—Tariq Akmal, Mike Trevisan, Steve Hollenhorst, and Scott Graves—for your careful reviews and the perspectives you provided. The educators who have made an impact on me over 25 years of schooling are too many to name here, but I think of your dedication often, and many of your lessons and the experiences you provided me will stay with me forever. I thank Fran Hermanson for demonstrating how to be a professional. I thank the staff at the Center for Assessment and those with whom I had the fortune to interact in NCME governance for your contributions to my professional development and making me feel like I have a place in the field. Finally, I thank all my fellow students for all of the conversations, intellectual and otherwise, the commiseration, and, most of all, the enduring friendships.

AN INVESTIGATION OF TEACHER EDUCATIONAL

MEASUREMENT LITERACY

Abstract

By Chad Martin Gotch, Ph.D.
Washington State University
August 2012

Chair: Brian F. French

The U.S. educational system is saturated with data on student achievement and performance that drive high-stakes decisions such as student promotion/retention, graduation, and teacher, principal, and school evaluation. To have confidence in these decisions, we need a workforce that is literate in assessment data—where they come from, what they can and cannot tell us. Many assessment data that are made publicly available come from standardized tests. Investigations of assessment literacy have traditionally emphasized classroom assessment. While the skills typically associated with this form of assessment, such as aligning in-class assessment to learning objectives and developing reliable and trustworthy grading methods (e.g., via rubrics), are worthy of study and cultivation, with the increasing visibility of standardized test data and its integration in to instruction and evaluation, a particular competence, namely measurement literacy, needs to receive more scholarly attention. Measurement literacy, as defined in this dissertation, concerns the ability to understand and work with the results of standardized tests.

This dissertation contains three manuscripts with the following purposes: 1) to provide a basis for establishing a collective memory in the area of empirical assessment literacy study, and

to identify gaps and inefficiencies in attention through a systematic review of the literature, 2) to examine the internal structure of a measure of educational measurement self-efficacy via factor analysis, and 3) to gather response process data to evaluate the extent to which a measure of educational measurement knowledge can support valid inferences of teacher understanding of measurement concepts. Results of this effort show we need improvements in measuring assessment literacy and assembling a cohesive community of scholars to build knowledge on the subject. Tentative support for the educational measurement literacy instrument was found. Some items appeared to function well while others need revision for the instrument to provide the most robust evidence for inferences about teachers' levels of measurement literacy. Future research should continue to evaluate evidence for the measurement literacy instrument; gather baseline information about the levels, antecedents, and associations of measurement literacy in the teaching workforce; and use this information to design effective professional development and communications around student test performance.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

We live in a society where numerical data bombards the public. News headlines report polling statistics for presidential candidates; prescription drug advertisements report probabilities of incurring side effects; car manufacturers tout crash safety star ratings. As citizens of the general public, we consume these reports, and to varying degrees use their information to make decisions. Specialized numerical data has also infiltrated many professions. From the farm to the courthouse to the automobile mechanic's shop to the office of professional sport scouts, making sense of numbers has become a required skill. Unfortunately, our best estimates suggest that many educated adults remain unable to function effectively in number-immersed environments (National Council on Education and the Disciplines, 2001).

This collective inability to operate in data-rich environments is particularly disconcerting in cases when these environments are also characterized as *high-stakes*. In contrast to *low-stakes* environments where immediate outcomes from decisions carry little influence over a person's wellbeing, high-stakes environments require one to interpret data to render decisions that carry important consequences (Madaus, 1988). High-stakes environments characterize the context within which a number of professional fields operate. For example, radiologists interpret medical imagery in concert with information from patient examinations and tests to diagnose illnesses and prescribe treatment. Military intelligence officers use information about terrain, weather, and location of an enemy to reduce the uncertainty of battle conditions. An appreciable degree of data literacy is required in order to arrive at accurate interpretations, and avoid serious negative outcomes such as misdiagnosis of terminal disease or sending soldiers into harm's way.

In the field of education teachers are asked to operate in a manner similar to these other professions. The role of assessment within public education has evolved, and the current structure produces many data on student performance. State and federal legislatures and offices of education use these data to make high-stakes decisions. The data often come from standardized test scores. In school systems today standardized tests are used for accountability purposes (e.g., adequate yearly progress (AYP) benchmarks), to render decisions about the educational path of students (e.g., high school graduation exams, placement into special education tracks), and to guide instruction (e.g., skill diagnostics, interim assessments). In many cases their use accompanies high stakes for students, teachers, and educational administrators alike. Concerns about the use of standardized tests data are primary inspirations for the work proposed in this dissertation.

After a brief review of the assessment context of K-12 public education in the United States, I will make a case for how this system could benefit from a workforce that is able to operate effectively in this context. I will then describe some key aspects that will need to be in place to strive for this end goal. The main purpose of this dissertation will be to articulate a specific kind of literacy, namely measurement literacy, that is a needed quality among the teaching workforce, and to provide validity evaluations for an educational measurement literacy instrument. This instrument will be discussed in detail below, and this chapter will conclude with discussion of how the proposed research fits within a broader context of research, policy, and practice.

The Role of Assessment in Public Education

Assessment for Accountability

Accountability demands are prevalent and persistent in the American public education system, particularly in reform movements of the past two decades. Standardized testing has shaped the national dialogue on educational reform, and the performance of U.S. schoolchildren on international standardized tests have been used to portray the U.S. educational system as losing ground in a global marketplace (e.g., Armario, 2010; MacInnis & Lambert, 2011). The No Child Left Behind Act of 2001 (NCLB; PL. 107-110) solidified on a national level a standards-based educational system, accompanied by mandates to assess reading and math proficiencies of all students. This federal school accountability law mandates all children in public schools between grades 3 and 8 have their achievement in math, reading, and writing assessed in a systematic and robust way. Across these grades are almost 22 million schoolchildren; another 1 million or more teachers are directly responsible for teaching the subject matter covered by large-scale accountability assessments (Department of Education, Institute of Education Sciences, 2010).

The extensiveness of accountability initiatives has accompanied intense pressures throughout school systems, understandably so, as consequences of unsatisfactory performance are game-changing. Students may be denied graduation, promotion, or access to beneficial learning resources/environments (Center on Education Policy, 2007; Darling-Hammond, 2007; Heubert & Hauser, 1999; Penfield, 2010). Teachers may have bonus pay withheld or lose their jobs all together (e.g., Greenhouse & Dillon, 2010; Sawchuk, 2011). School administrators may lose their jobs as they are replaced in "turnarounds" (U.S. Department of Education, 2010, p. 12). Teachers can be stigmatized by being associated with low-performing schools (specifically

ones undergoing mandated restructuring), and face challenges finding employment at another school (Blume, 2011). More than a dozen states have developed and are beginning to implement teacher evaluation systems—to be used for actions such as promotion, salary bonuses, and termination of employment—that incorporate use of student test scores (Vaznis, 2011). Potentially in response to these consequences (Herold, 2011), widespread evidence of cheating has been found in major school systems (Cizek, 2011).

Despite these pressures, however, educational reform efforts have not reduced the importance of standardized assessment. The Council of Chief State School Officers (CCSSO) and National Governors Association (NGA) pushed for a set of Common Core Standards on which to measure students' growth and proficiency. Such an action signals strong support by high-ranking educators to maintain an educational system based on strong standards against which students can be evaluated. The federal Race to the Top initiative encouraged states to work together to develop high-quality assessments utilizing the Common Core Standards (Council of Chief State School Officers & National Governors Association Center for Best Practices, 2010). These partnerships, SMARTER Balanced Assessment Consortium (SBAC) and Partnership for the Assessment of Readiness for College and Careers (PARCC), are comprised of a combined 45 states with some states joined up with both consortia. The consortia proposals utilize innovative summative, interim, and formative assessments to build a system of benchmarking and feedback aimed at increasing the number of students who graduate high school ready for entry into college or a career. Such a strategy is often called a balanced assessment system. Early indications from the work of the consortia suggest assessment will become more integrated into day-to-day activities (Gewertz, 2011b). Accountability assessment in the form of standardized tests shows no signs of diminishing.

The consortia are looking to leverage technological advances in testing (e.g., computer adaptive test environments) to provide frequent, high-quality feedback and measures of growth in a fiscally sustainable way. Prior conceptions of balanced assessment systems have highlighted the need for professional development to assist educators with the transition to this new methodology (e.g., Bennett & Gitomer, 2008). The assessment consortia have recognized this need, and have substantially bulked up their financial commitment to the development of curriculum resources, instructional materials, and professional development workshops (Gewertz, 2011a).

Assessment to Identify Students for Special Services

Another use of assessment that perhaps does not garner the media attention of accountability-oriented assessment but is impactful nonetheless is the use of test results to determine a student's eligibility for special services. Such special services may include development of an individualized education program (IEP) or provision of a learning environment for gifted students, in other words, services that are essential to meeting the individual's needs (Shapiro, 2011). Relatedly, assessment results also may be used to track students into specific learning pathways (Shapiro). In all of these cases, the fate of a child's educational career hangs on his or her performance on some kind of assessment, often a standardized measure. Consequences of misidentification may not be as immediately apparent as grade promotion denial, for example, but are serious and potentially more impactful over time.

There also are considerations to make at the aggregate level. Special services are generally a scarce resource. As such, only a small proportion of the population may qualify to receive these resources. Allocations for these services are based to some extent on perceived need. Gross misinterpretation of testing and test scores used to identify perceived need can result

in either not adequately funding special services—denying optimum learning opportunities in the process—or using finite public dollars ineffectively. With the correct interpretation of performance, the student stands to benefit from the optimal learning resources available, and the public benefits from a greater collective good.

Assessment to Guide Instruction

Finally, there is assessment that is used to guide instruction. Within this type of assessment lie formative efforts, where the teacher and student use assessment performance to target learning activities. Brookhart (2007, p. 43) states formative assessment "gives teachers information for instructional decisions and gives pupils information for improvement." Interest in formative assessment has grown recently in response to frustrations that high-stakes tests used for accountability purposes have driven much of what goes on in the schools, but not provided enough support to guide teaching and learning. Because of its direct ties to the act of instruction and consideration for the needs of individual students this form of assessment could be considered closest to the interest of teachers.

The character of formative assessment has blurred in recent years, however. Recent changes in demand for formative assessments and in the growing accountability-orientation of assessment systems have created some murkiness regarding what counts as formative in nature. There has been a rise in interim assessments, which while often marketed as formative in nature, may actually constitute a fundamentally different kind of assessment and therefore cannot draw from the substantial literature base that praises the instructional benefits of formative assessment (Perie, Marion, Gong, & Wurtzel, 2007). Counter to this claim, there is some evidence that the act of completing a test can enhance retrieval processes and lead to better test performance (Karpicke & Blunt, 2011). This evidence represents a testing effect that has been documented for

decades, if not longer (Roediger, III & Karpicke, 2006). Now that we have reviewed common purposes for assessment, let us consider what we may reap from having a teaching workforce that is well-versed across these purposes.

Benefits of a Measurement Literate Teaching Workforce

A theme across the assessment roles described above is the pervasiveness of standardized testing. Teachers' abilities to navigate a sea of test data impact their students as well as their own employment futures. Consider the impact standardized assessments have had on teachers and their classrooms and the potential that lies in building teachers' capacities to understand and work with test results. We will refer, heretofore, such capacity as *measurement literacy*. Measurement literate teachers can understand the performance of their students. In other words, the test results could provide the most constructive feedback to teachers if they develop a nuanced grasp of what can be gleaned from the results. Study 1 will investigate this quality in greater detail.

In a thorough review, Black and Wiliam (1998) catalogue many primary studies and meta-analyses provides substantial evidence of the positive impact well-designed assessment practices can have on student learning. In addition to a focus on academic performance, Stiggins (2002, 2007) has argued for consideration of the impact assessment has on the emotions of students, their self-evaluation as scholars, and their motivation to learn. Brookhart (1997) advanced a theoretical framework for the relationship of classroom assessment practices to student motivation and achievement. Empirical scrutiny of such a relationship revealed significant relationships between student academic performance and teacher assessment practices (Rodriguez, 2004). These findings, combined with the work of Black and Wiliam and the

arguments advanced by scholars suggest that the skill with which a teacher can manipulate assessment-related activities can impact student effort and achievement.

Present shifts in the assessment landscape indicate literacy related to educational measurement could be gaining importance. Recall the balanced assessment systems being developed by SBAC and PARCC seek to integrate assessment and instruction. In these systems, classroom activities begin to take on both instructional and evaluational purposes. With the introduction of interim assessments, some of which may be likely scored in a similar way to the accountability-oriented tests given once or twice a year, the ability to make sense of scale scores, standard errors, reliabilities, becomes even more essential. Quickly-scored interim assessments, in combination with heightened levels of measurement literacy, could bring traditionally summative assessments, ones that previously have been seen as apart from the instructional process, an impediment even, into the sphere of instructional decision-making during the schoolyear, when the teacher may still act upon the present class of students. Measurement literacy could empower teachers to use this information, interpret it with the proper caveats, and effectively target instruction.

Effectively targeted instruction benefits the student directly and the teacher indirectly (via potentially improved student performance on metrics used for teacher evaluation). There is a benefit, however, that teachers may experience directly from attaining high levels of measurement literacy. That benefit is the potential to become active and engaged citizens in the educational reform arena. An improved literacy regarding assessment data can be liberating and empowering for teachers, engendering more democratic educational policy developments (Freire, 1973/1998), which could be particularly beneficial considering cases where students'

tests are going to be used to evaluate teachers. It makes sense to have teachers be able to understand what's being used to evaluate them.

National teachers unions have released formal statements on the use of test scores to evaluate teachers. The National Education Association has endorsed teacher evaluation systems that use "developmentally appropriate, scientifically valid and reliable for the purpose of measuring both student learning and teacher's performance", and the American Federation of Teachers, similarly, advocated the incorporation of "valid assessments" in teacher evaluations (Otterman, 2011). These entities have substantial potential impact on classrooms and policy alike. These statements, however, beg the question of whether or not the teacher union members can identify what would make an assessment's scores valid.

Building and Assessing Measurement Literacy

To develop a body of teachers that is functionally measurement literate, we need to give thorough consideration to how we conceive of literacy. Literacy is more than possession of a collection of skills. Fundamental knowledge of facts is essential, but does not encompass the totality of literacy. Well-developed vocabularies are foundational for literacy but insufficient for defining literacy on their own (Lonigan, Schatschneider, & Westberg, 2008, p. 75). One must consider not only technical skill but also what the individual does with the technical skills at his disposal, how frequently those skills are called upon, and to what ends they are used (Cremin, 1988). Therefore, there is a behavioral component, a kind of internalization of competencies and an orientation to make using assessment results a habit.

Before any defensible claims can be made about the current state of measurement literacy among education professionals and any changes in this state, we need to develop a solid measure of the phenomenon in which we are interested. While anecdotal evidence is easy to come by, and

calls for improving teachers' skills are easy to find support for, it is empirical data that will drive the science behind training programs and policy efforts. To collect such empirical data, we need an instrument that can measure teacher literacy in working with test results, and we need an instrument that can capture the multifaceted nature of literacy—both technical vocabulary and inclination toward certain behaviors. To date, no such instrument has been published or used in published, peer-reviewed studies.

Given the evolved nature of the assessment results available to teachers, students, parents, and the public, this instrument needs to be able to capture the understanding teachers have for quantitative test results. Developing effective in-class assessments, appropriate scoring and grading methods, and an ability to identify unethical assessment practices are essential elements of any teacher's repertoire of job skills, but the pervasiveness of standardized testing and the visibility of the results of such assessment necessitate the development of an instrument tailored to this specific kind of test result data. For this purpose the Teacher Educational Measurement Literacy Scales (TEMLS) were developed.

Instrument

The TEMLS were developed to match the content of student score reports from several U.S. states, Canadian provinces, and commercial vendors, supplemented with items that addressed fundamental educational measurement concepts such as validity, reliability, and the aims of different kinds of assessments (e.g., formative, summative, norm-referenced, criterion-referenced). Concepts addressed by the TEMLS are well-represented in measurement and assessment texts that would be used in a teacher training program (e.g., Linn & Miller, 2005; McMillan, 2007; Popham, 2005; Woolfolk, 1995). Developed items were reviewed by an expert

item writer and a panel of local school district personnel working on assessment and score reporting issues.

Measurement knowledge scale. This scale consists of 20 multiple choice (4 options) items covering issues such as interpretation of standardized scores; scores in relation to one another within a student, across students, and across schools; and proficiency level interrelation. Example question stems included: (a) *Evan, a third-grader, obtained a percentile rank of 90 on a standardized reading assessment. This indicates Evan…*, (b) *Proficiency exams are primarily used for determining if…*, and (c) *If Mrs. B. wanted to know Elise's strengths and weaknesses on certain reading skills what type of assessment would be most helpful?*

Measurement self-efficacy scale. The inclusion of the self-efficacy scale begins to address the behavioral and referent aspect of literacy described above. This concept aligns very well with discussions around the nature of literacy as "perceived self-efficacy is concerned not with the number of skills you have, but with what you believe you can do with what you have under a variety of circumstances (Bandura, 1997, p. 37)". Self-efficacy therefore, addresses the "difference between possessing subskills and being able to integrate them into appropriate courses of action and to execute them well under difficult circumstances". Therefore, the self-efficacy scale serves as a necessary complement to the knowledge scale to more fully capture the notion of literacy in the measurement domain.

This scale consists of 21 items assessing a teacher's judgment of his or her capabilities to use test score information. Items were modeled after Bandura's (2006) guidelines for developing self-efficacy assessments. The items employed the stem, "How well do you believe you can...", and were rated on a 7-point scale (1 = Not at all well, 7 = Very well). Example items included (a) Explain a scale score to a student's parent, (b) Understand validity information presented in a test

manual, and (c) Identify whether or not a student's test score meets a specified standard. Rating scale construction diverted from the recommended 11-point scale using 0 and 100 as endpoints because 7-point scales have been shown to maximize reliability and validity of responses (Lozano, García-Cueto, & Muñiz, 2008).

Research Endeavors

The primary endeavor here is building a measurement instrument. As a part of building that instrument there are some necessary steps to complete. To build a solid foundation for future work based on this line of research, we must ensure this measurement instrument can provide grounds for valid inferences about teacher measurement literacy. Just as ensuring teachers draw accurate inferences from test results is of high priority, so is making sure that the inferences we draw about teachers' abilities are equally accurate. Therefore the research presented here comprises a validity investigation of the TEMLS content and scores coming from TEMLS responses.

Validity

Validity concerns the appropriateness of inferences drawn from data (Sireci, 2009). These inferences typically center on a construct of interest. In education and the social sciences, such constructs of interest often reside in the latent world. That is, the things we are concerned about (e.g., reading ability, confidence, anxiety) cannot be observed directly. Instead, we rely on measures that we believe capture how these latent constructs manifest themselves in observable behavior (e.g., decoding phonemes) or involuntary responses (e.g., heart rate). When we use these manifestations, we implicitly assume they accurately represent the phenomenon that is actually of interest. Support for this assumption is a necessary requirement for our inferences to be valid.

Messick (1989) outlined five common sources of validity evidence, which have guided much validation work over the last 20 years. These sources are a) test content, b) internal structure, c) response processes, d) associations with other measures, and e) consequences of testing. Each of these sources should inform the user of a test instrument to what extent data from that instrument can support inferences about the construct of interest. For example, if an instrument is purported to measure distinct facets of a construct, investigation of the internal structure of the data can inform the user the extent to which these facets are reflected and distinct from one another. Investigation of test content can reveal the extent to which the construct of interest is well covered by the test items and therefore to what extent inferences from the data align with inferences about the construct.

As validity theory has developed, people have taken Messick's conceptual framework and applied an action component, moving the discussion from validity to validation. The most common contemporary notion of validity places the act of validation in an argumentation framework (Kane, 2006). This framework involves a chain of interpretative arguments. Each interpretive argument moves from an observation to a claim. One interpretive argument's claim becomes the observation in a subsequent interpretive argument. Collectively, this chain provides contextualized explanations of test scores (Zumbo, 2009).

To bring the work of validation to the TEMLS, first, we need to refine our understanding of the construct of interest. What exactly is measurement literacy and how is it different from other, similar constructs (e.g., assessment literacy) for which measurement instruments already exist? Once that process has finished, we need to begin to build evidence that the instrument is functioning well and gathering an accurate account of the intended conception of measurement literacy. Such efforts represent the work embodied in this dissertation.

Outcomes of the Proposed Research

This instrument has been used to collect baseline data on a sample of elementary school teachers. The data gathered from administration of the instrument, however, and the associated inferences drawn from them can only be supported if the instrument has gone through rigorous validation. Therefore, three separate but associated objectives are proposed for the present research effort:

- The first objective will be to trace a history of recent assessment literacy research, particularly endeavors that required measurement of such literacy. This investigation will reveal the extent to which gains have been made in the study of assessment literacy since the publication of two major works in the area.
- A second objective will be to investigate the internal structure of the self-efficacy component of the TEMLS. This investigation will reveal whether measurement self-efficacy can be considered a unitary or multifaceted concept. This analysis will complement analysis that has been conducted on the knowledge scale component of the instrument (Gotch & French, 2011).
- A third objective will be to examine the response processes that individuals go through when completing the knowledge portion of the TEMLS. From investigation of these processes evidence of the instrument's ability to accurately reflect knowledge of measurement concepts will be provided.

By achieving the three objectives through the three studies proposed, evidence for weighing the merit of claims of teacher educational measurement literacy based on TEMLS data will be provided. Achieving the objectives will also illuminate potential revisions to the instrument. The studies will work in concert to guide these revisions. Results from each study

will be synthesized in the conclusion section, with specific statements addressing claims regarding what the TEMLS can and cannot support.

Significance of the Research

Standardized testing impacts millions of children and educational professionals across the United States. Stakes are high for those directly involved in the educational process, but the effects of a large-scale accountability system reach beyond those with direct involvement. We rely on our schools to prepare our next generation of citizenry. To progress as a society we need quality in our educational system. In the current climate of accountability, a quality educational system depends on a professional body that is literate in assessment and can act appropriately in response to test results. Establishing teacher competence for working with test data is imperative given such large numbers of students enrolled in the schools and the high-stakes decisions being driven by test results. Researchers and those responsible for teacher training have devoted a wealth of writing to the impact of testing on students and teachers, and made multiple formal calls for competencies teachers and other educational professionals need to have with regard to student assessment. This attention has set the stage for developing measures of such competencies, and using empirically-derived data to develop and evaluate teacher training initiatives.

Structure of the dissertation

The contents of this dissertation are presented as three independent manuscripts, followed by a concluding chapter. Each manuscript is intended to stand on its own, resulting in some redundancy across the entire volume. References and appendices cited in this work are consolidated into a single section. Tables and figures appear with their relevant manuscript.

CHAPTER TWO:

A SYSTEMATIC REVIEW OF EMPIRICAL ASSESSMENT LITERACY SCHOLARSHIP

The act of testing has been around for millennia, spanning back to ancient Chinese and Greek cultures (Anastasi, 1993; Goodenough, 1949). Through the years, testing programs have fulfilled a variety of needs such as evaluating fit for civil service, assessing mastery of skills, identifying individuals with particularly advanced or delayed development, and licensing individuals for employment practice, to name a few among many (Aiken, 1991; Anastasi & Urbina, 1997). The outputs of tests, in the form of examinee responses, are themselves meaningless; they require someone to ascribe meaning to them within the broader context of the purposes of the testing program (Crotty, 1998, p. 43). One could reason, therefore, that interpreting test performance is as central to the purpose of testing as the administration of the test itself. Having a body of professionals who were literate in the assessments they conducted has always been essential.

Just over 20 years ago, two publications set the stage for a formal area of study around the ability of teachers and other educational professionals to utilize the process of assessment in its most effective and ethical way. The first of these publications was the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990). A collaborative effort between measurement specialists and teacher representatives, the Standards generally set forth expectations that teachers possess abilities to properly select, develop, and carry out a variety of appropriate assessment methods using both self-produced and externally-

produced measures; use assessment results to plan instruction and make decisions about students; and communicate assessment results to stakeholder parties (e.g., students, parents).

Following soon after the *Standards* was a piece in *Phi Delta Kappan* by Richard Stiggins (1991) in which he coined the term *assessment literacy*, and described assessment literate teachers. Such teachers "have a basic understanding of the meaning of high- and low-quality assessment and are able to apply that knowledge to various measures of student achievement (p. 535)". Stiggins further expounded on this core description by adding that assessment literates ask what a given assessment tells us about valued achievement outcomes and how students will be affected by the assessment. Assessment literates understand the values of employing assessment methods that precisely and thoroughly target a desired outcome. They clearly communicate these methods, and have a built-in awareness of factors that can interfere with the assessment being carried out properly and providing meaningful results. By providing this tangible, unified vision of assessment literacy, Stiggins's article marked a milestone in a line of research extending back several decades. This line will be traced briefly in the next section.

The perceived importance of teacher competence in understanding, employing, and making sense of student performance remains strong, as prominent scholars continue to advocate and conceptualize requisite skillsets and orientations teachers should possess in relation to assessment (e.g., Brookhart, 2011a; Popham, 2011). Now that two decades have passed, and the shape and influence of our assessment systems have changed in response to federal accountability mandates, it is time to reflect on the concept of assessment literacy and evaluate how far we have come as a research community. To what extent has the study of assessment literacy matured and what knowledge have we built? This article will trace a brief history of the study of assessment literacy, and then focus heavily on empirical work that has been produced in

the field in the last 20 years. The aim of this study is to provide a basis for establishing a collective memory (McNeill, 1985) in the area and identify gaps and inefficiencies in attention.

Arriving at a formal terminology

Formal investigation of teachers' competencies for interpreting test performance can be traced back at least half a century, to the work of Noll (1955), who looked for educational measurement coursework as a part of state requirements for teacher certification. He found few states held such requirements despite apparent widespread availability of such courses in the universities (as reported in Schafer & Lissitz, 1987). The topic was picked up again the late 1960s to early 1970s. Mayo (1967, 1970), Roeder (1972, 1973), and Goslin (1967) all continued the investigation of measurement coursework in the background of classroom teachers. Results remained consistent with Noll's findings. Goslin also examined uses of standardized tests, teacher opinions of tests, and various roles teachers could adopt in relation to tests and testing programs. Goehring (1973) reviewed educational testing textbooks and conducted a teacher task analysis to provide recommendations for the types of competencies pre-service coursework should seek to build in future teachers. He concluded that training in processes (e.g., construction, application, interpretation, and evaluation of tests) was more important than training in statistics.

Through the late 1970s and into the mid-1980s, methods of investigating teacher assessment competence diversified to include administrators' perceptions (Hills, 1977) and analysis of teacher-developed tests (Flemming & Chambers, 1984). During this time, as reflected in the proceedings of the 1980 National Institute of Education Conference of Test Use, researchers carried on Goslin's initial investigations by further examining teacher attitudes toward testing, classroom assessment practices, and how teachers used test results (Rudner,

1980). This body of work generally indicated that teachers had a more positive opinion of tests than expected by the researchers, and could recognize specific values of using tests and being well-versed in testing concepts, but that teachers were using standardized tests more for identification of student growth and strengths than for instructional planning. The early 1980s also saw research broaden to include school counselors as a population of interest in the assessment literacy domain (e.g., Daniels & Altekruse, 1982). Counselors and testing had been linked for a long time (see Schafer & Mufson, 1993), but focus had primarily been on what tests could tell the counselor rather than what information the counselor was able to glean from the test.

Arlen Gullickson carried the torch for research connecting teachers and assessment through the mid-1980s, publishing work on teachers' attitudes toward testing and use of tests (Gullickson, 1984), student evaluation techniques (Gullickson, 1985), and teachers' needs in educational measurement (Gullickson, 1986; Gullickson & Hopkins, 1987). In general, he found teachers to be open to the information tests could provide, but more interested in practical concerns regarding testing and how they could gather information through means other than tests. He also found teachers' assessment knowledge to lag behind what they would need in order to use tests most effectively, and suggested a need for strategic changes in teacher training in educational measurement concepts. The mid-1980s also saw Stiggins's early attempts to establish the importance of classroom assessment to educational effectiveness (Stiggins & Bridgeford, 1985; Stiggins, Conklin, & Bridgeford, 1986). Schafer and Lissitz (1987) conducted an extensive review of how measurement was incorporated into university teacher preparation programs. Consistent with the work conducted 20 and 30 years earlier, they found teachers were not exposed to much measurement instruction in their pre-service training.

While no clear emphasis on educational measurement or assessment literacy existed among the states at the time (Stiggins & Conklin, 1988), professional organizations were undertaking large-scale efforts to highlight the need for competencies in assessment. At the 1989 National Council on Measurement in Education (NCME) Annual Meeting, members debated what types of assessment training opportunities were most essential for classroom teachers (O'Sullivan & Chalnick, 1991), reflecting the work that had begun on developing the Standards. A collaborative effort, NCME worked with the major teachers' unions to produce this landmark publication. A formal declaration of expectations for teacher competence in assessment had been set. Following soon after, Stiggins authored his Phi Delta Kappan piece on assessment literacy. The formal declaration of expectations had now been given formal terminology. These two works galvanized the efforts of prior decades that focused on the training, understanding, and use of tests and other assessments, and set the stage for a formal area of study to be further conceptualized and explored.

The scholarship of assessment literacy

The focus of this review is on empirical work in the area of assessment literacy. Advocation is aplenty as a dedicated few have maintained consistent calls or engaged in work to improve assessment literacy in periodicals (Brookhart, 1997, 1999, 2003, 2011a; Popham, 2003, 2006, 2009, 2011) as well as practitioner-oriented outlets (e.g., the Kansas State Department of Education's Assessment Literacy Project). Stiggins continues to publish on the matter, particularly in relation to formative forms of assessment (e.g., Chappuis, Stiggins, Chappuis, & Arter, 2012; Stiggins & Chappuis, 2006, 2008), and heads Pearson's Assessment Training Institute. But to what extent has the scientific research community picked up on these calls and contributed to an empirical knowledge base? In a world of finite resources and competing

demands for research and educational funding, a strong case needs to be made for the measurement and usefulness of assessment literacy. Ultimately evidence needs to be presented alongside anecdote and impassioned pleas.

With these considerations in mind, the present research advanced the following questions:

1) What is the body of empirical work from the last two decades that directly addresses the study of teacher assessment literacy?

2) What have scholars done to investigate assessment literacy in terms of the extent to which it exists among the teaching workforce, the impact of initiatives designed to increase teacher competencies, its relation to other variables and characteristics?

Within these two guiding questions, effort was taken to characterize the outlets chosen for disseminating knowledge on this topic and the form in which the empirical work appears (e.g., scholarly journals, reports, conference papers). Who has contributed to the knowledge base and from what settings? Findings from the research literature base were summarized. This paper concludes with a discussion of the strengths and shortcomings of existing research and a call for a more considered, thorough, and consistent system of investigation of assessment literacy.

Method

This study took the form of a systematic review of the literature. Research published between 1991 and 2011 was searched for scholarly works readily available to researchers that answered the research questions outlined above. The ERIC, PsychInfo, JSTOR, Education Full Text, and ProQuest Dissertations and Theses databases were searched using the terms *assessment literacy*, *measurement literacy*, the combination of *educational assessment* or *student*

assessment with *teacher competenc**, and the combination of *educational assessment* or *student assessment* with *teacher training* or *teacher preparation*. The search terms were kept narrow because the central interest of the review of literature was assessment literacy as defined by Stiggins. Certainly other resources that discuss the uses of assessments or teacher training in general would give tangential attention to assessment literacy. However, those resources would be unlikely to contribute to the story of assessment literacy as a formally defined concept.

The term *measurement literacy* was included to reflect recent discussions in the field. For example, during his tenure as President of NCME, Terry Ackerman, in an interview published in the organization's newsletter, discussed a growing demand to make teachers and administrators "measurement literate" (Barry, 2009, p. 10). Measurement literacy represents a subset of assessment literacy, aimed specifically at the use and interpretation of standardized tests.

As a final check, Google Scholar was referenced for any additional published work that cited the 1991 Stiggins article or his 1995 follow-up of a similar focus. This formal search returned 1,097 articles, books, book chapters, reports, conference papers, dissertations, theses, reviews, unpublished manuscripts, and errati. For the sake of clarity, all of these works will be referred to as *studies* through the remainder of this document. Abstracts, and where necessary to obtain a clearer understanding of content coverage, whole copies, of these studies were reviewed to evaluate the extent each possessed as a central focus a) the examination of baseline levels of assessment literacy, b) evaluation of the outcome of an intervention aimed at improving assessment literacy, or c) the study of assessment literacy in relationship to other variables. If a given study held one of these aims, and was carried out in the P-12 U.S. setting or an international equivalent, it was retained for further review. Studies were not retained if they promoted a general need for assessment literacy, outlined a conception of assessment literacy,

proposed a framework for the cultivation of assessment literacy, provided instruction on assessment, investigated teacher training and certification requirements, or held any other similar aims without including a direct examination (i.e., measurement) of assessment literacy to fulfill one of the focuses listed above. Studies were also not retained if they focused on the assessment literacy among instructors in the higher education setting, a population other than teachers (e.g., principals, counselors) to the exclusion of teachers, or only on tangential constructs like attitudes toward testing, beliefs about types of assessments, or assessment practices (without an evaluation of such practices). Studies of self-perceptions of assessment literacy were, however, included for review. While the interest was in empirical investigation of assessment literacy, inclusion was not limited to a particular research design (e.g., pre-test/post-test with large samples) or operation of the construct (e.g., test of assessment knowledge).

Studies that met the stated inclusion criteria were then obtained and the reference sections of these works were checked for any other published material that might be suitable for inclusion in the review. Additionally, obtained studies were checked for redundancy. For example, in some cases the same study was published in two different periodicals or a study first presented in a dissertation or conference paper was later published in a journal. In the latter case, the journal article was retained as the case of record. In the former, a judgment was made regarding which journal represented the highest level of publication (e.g., national vs. regional). After these checks, a final set of 65 published works was identified for review—36 articles, 17 dissertations, 10 conference papers, 1 report, and 1 book chapter. Appendix A presents basic information for the studies under investigation in this paper. Certainly other studies have been conducted on the topic, and presented at national and regional conferences, but if the written outputs of the studies

are not readily accessible to researchers, these studies hold little potential to add to the knowledge base.

The analytical phase of the study involved a careful reading of included studies for their authorship, purpose, study sample, measurement of assessment literacy and other variables, and findings. Themes in the literature were established by identifying commonalities in these attributes across studies. In several cases a single study was unique in such attributes (e.g., a specific form of measurement), and in other cases contrasting findings arose. Such cases are described in further detail in the next section.

Results

Defining assessment literacy in practice

Not surprisingly, given the search methodology, Stiggins (1991, 1995) was by far called upon the most for explicit conceptual definitions of assessment literacy. What ultimately is more important in the context of this study, however, is how assessment literacy (both explicit and implicit investigations of it) has been operationalized. That action is what distinguishes the literature under investigation in this study from all of the words that have been devoted to conceptualizing assessment literacy. Empiricism requires operation. One study from early in the timeframe of years considered for this review established a measure of assessment literacy that served as the basis for several subsequent studies. That study was the one conducted by Plake, Impara, and Fager (1993), and the instrument was the *Teacher Assessment Literacy Questionnaire* (TALQ). The TALQ was the first to be administered to a national sample of participants and to receive prominent attention in the measurement community. It was based on the *Standards*, and targeted a series of multiple-choice questions toward each of the declared competencies. This instrument was used and adapted in later work (Benson, 1997; P. Chen,

2005; O'Sullivan & Johnson, 1993; Scribner-Maclean, 1999), sometimes going by the name *Assessment Literacy Inventory* (Mertler, 2009; Mertler & Campbell, 2005) or *Classroom Assessment Literacy Inventory* (Mertler, 2003). It even was translated for use in Arabic (Alkharusi, 2011a).

Other psychometric instruments, some based directly on the *Standards* (Arce-Ferrer, Cab, & Cisneros-Cohernour, 2001; Barr, 1993; Braney, 2011; Zhang, 1996; Zhang & Burry-Stock, 1995, 2003) and some not (Daniel & King, 1998; Fan, Wang, & Wang, 2011; Wang, Wang, & Huang, 2008), were also employed in the study of assessment literacy. Most of these instruments took the form of a test of assessment knowledge, but others relied on self-report of perceived competence (Arce-Ferrer et al., 2001; Hambrick-Dixon, 1999; Kershaw, 1993; Scott, Webber, Aitken, & Lupart, 2011; Volante & Fazio, 2007). The soundness of the various forms of measures of assessment literacy was assessed to varying degrees, at least as can be gathered from published work. Scale descriptive and reliability calculations typically characterized the extent of psychometric analysis reported on the instruments, though a few exceptions to that rule existed. Alkharusi (2011a) carried out a series of additional analyses (e.g., factor analysis, associations with a criterion) to judge the suitability of a translated version of the TALQ for use in the Omani population. Zhang (1996) submitted the *Assessment Practices Inventory* to principal components analysis and Rasch modeling to gain insight into the functioning of items on this instrument.

To investigate teachers' competencies for assessment through means other than scores on an instrument was quite common. Over one-third of the studies included such alternative data sources. Teacher work samples were a common mode of collecting information to evaluate assessment literacy levels (Arter, 2001; Bangert & Kelting-Gibson, 2006; Buck, Trauth-Nare, & Kaftan, 2010; Campbell & Evans, 2000; McMorris & Boothroyd, 1993; Schmitt, 2007; Siegel &

Wissehr, 2011). Interviews and focus groups were also common (Borko, 1997; Bruce, 2004; Lomax, 1996; Scott et al., 2011). Reflective journals, philosophy essays, and other written passages about assessment (Lomax, 1996; Maclellan, 2004; Siegel & Wissehr, 2011) as well as observations of practice (Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Forbes, 2007; Scribner-Maclean, 1999) were also employed by researchers. There was no evidence provided among the studies to suggest one study's methodology built on the methodologies of studies by other authors in the review.

The treatment of assessment literacy was both global and domain specific. Some studies looked at assessment literacy among teachers across the spectrum of content responsibilities, while others employed samples within a specific instructional domain (physical education, Williams & Rink, 2003; vocational education, Kershaw, 1993; science, Buck et al., 2010; Mazzie, 2008; math, Benson, 1997; science and math, McMorris & Boothroyd, 1993). Assessment literacy was defined and investigated for pre-service teachers or teachers returning to school for an advanced degree in 22 studies. The remainder of the studies focused on in-service teachers. Some studies carrying out separate analyses for each group (Alkharusi, Kazem, & Al-Musawai, 2011; Mertler, 2003, 2004). Both elementary and secondary teachers were studied with about equivalent emphasis. Focus on teachers in the middle grades was granted less emphasis, but did exist (Alkharusi, 2011b; McMorris & Boothroyd, 1993).

Assessment literacy in the teaching workforce

Overall evaluations of assessment literacy among teachers have not painted a rosy picture. Basic evaluations have described low levels of assessment literacy (Barr, 1993; Greenstein, 2004) and gaps in knowledge (Edman, Gilbreth, & Wynn, 2010; Maclellan, 2004). Teachers have been characterized as having difficulty with common assessment responsibilities

(Lomax, 1996) and naïve conceptions of the purpose of assessment (Maclellan, 2004) and of validity and reliability (Mertler, 2000). Evaluation of their work has shown they can produce rubrics of average quality (Schmitt, 2007), but do not demonstrate best practices (Lingard, Mills, & Hayes, 2006) or clear connections between instruction and assessment (Greenstein, 2004). Teachers' perceptions of their own assessment competencies could be generally high (Arce-Ferrer et al., 2001), but they still acknowledged certain assessment practices (e.g., test construction) can be complex, and doubted many teachers have the requisite skills (Scott et al., 2011).

At a finer grain of analysis, choosing assessment methods (Kershaw, 1993; Mertler, 2003, 2004; Plake et al., 1993) was found to be a relative strength of teachers, across studies. The specifics of item analysis (Kershaw, 1993; McMorris & Boothroyd, 1993) and developing grading procedures (Mertler, 2003, 2004) were consistently found to be more challenging. However, a more complex profile of teachers appeared across most of the analyses that investigated specific skills sets and behaviors. Variability among teachers and their particular assessment competencies can be large (Bruce, 2004; Greenstein, 2004). Administering assessments has been found to be a relative strength among teachers (Kershaw, 1993; Mertler, 2003, 2004; Plake et al., 1993), but also a relative shortcoming (Arce-Ferrer et al., 2001; VanLeirsburg & Johns, 1991). The same goes for competency in communicating assessment results (Arce-Ferrer et al.; Kershaw, 1993 vs. Mertler, 2003; Plake et al., 1993) and item writing (Chirchir, 1995; McMorris & Boothroyd, 1993). Gauging the degree to which teachers are successful with interpreting test scores is complicated by how that skill is measured across studies. In studies following the reporting scheme employed by Plake et al., interpretation of scores is reported within the same domain as administering and scoring assessments. In these

studies, that domain was a relative strength of teachers (Mertler, 2003, 2004; Plake et al., 1993). When interpretation of assessment results was studied and reported on with greater isolation, results were mixed. Interpreting standardized test scores such as percentiles and grade equivalents was a positive outcome for teachers in one case (Daniel & King, 1998), but an area of difficulty in other studies (Arce-Ferrer et al., 2001; Kershaw, 1993). Furthermore, understanding the broader context of test score properties, including standard error (McMorris & Boothroyd, 1993) and psychometric and statistical issues related to tests (Daniel & King, 1998), was found to be difficult for teachers at the aggregate.

Some findings within the area can only be attributed to a single study. Competency strengths for teachers were found in providing fair warning and an explanation of test purposes (VanLeirsburg & Johns, 1991), use of rubrics (Williams & Rink, 2003), obtaining validity evidence (Arce-Ferrer et al., 2001), and grading of traditional pencil-and-paper items (Kershaw, 1993). On the other side of the ledger, Scott et al. (2011) provided the only detailed look at understanding of interrelationships among policies, practices, and terminology, finding perceived shortcomings among teachers. Kershaw (1993) provided the lone study in which teachers were found to have low competency for scoring essays and compiling portfolios.

Outcomes from assessment literacy training

Published studies generally reported positive effects of professional development programs tailored to improving assessment literacy and state or provincial policies affecting the role of testing or use of assessment results. Such initiatives were shown to have improved teachers' abilities to develop quality assessments and engage in appropriate assessment practices (Bandalos, 2004; Borko, 1997; Borko et al., 1997; Forbes, 2007; Koh, 2011; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Mertler, 2009; Sato, Chung, & Darling-Hammond, 2008; Scribner-

Maclean, 1999; Vanden Berk, 2005). Web-based instructional modules were shown to produce at least tentative improvements in teachers' understandings of statistical concepts underlying test items (Fan et al., 2011; Wang et al., 2008; Zwick et al., 2008). Improvements also were found in competencies related to communicating results (Mertler, 2009) and recognizing unethical, illegal, or otherwise inappropriate assessment methods (Scribner-Maclean, 1999). Gains in assessment literacy could be generic (Mazzie, 2008) or very specific, as in Broodhead's (1991) study of teachers' abilities to diagnose levels of cognitive and affective development in high school students' dialogue.

Findings from investigations of regular coursework focused on preparing students in assessment produced mixed results. Portfolio-based (Taylor & Nolen, 1995) and online learning environments (Alkharusi, Kazem, & Al-Musawai, 2010) showed better outcomes than traditional modes of instruction. Contrary to the narrative of improved assessment literacy, Campbell and Evans (2000) found preservice teachers did not address fundamental validity and reliability criteria even after training. Volante and Fazio (2007) found teachers made little distinction between formative and summative assessment, and reported consistently low ratings of self-efficacy across 4 years of a teacher training program.

Associations between assessment literacy and other variables of interest

A common theme among studies conducted over the last 20 years was an attempt to reveal associations between assessment literacy and various other measureable qualities. Such exploration is critical to the study of assessment literacy in order to situate the construct within broader models of teacher training and educational reform. Assessment literacy was commonly investigated in relation to basic teacher demographics with mixed results. Advanced education (e.g., obtaining a graduate degree) was found to be related to higher assessment literacy (Hoover,

2009; King, 2010), but certification was not (Kershaw, 1993; King, 2010). Consistent with the evaluation of assessment trainings, outlined above, teachers with measurement training in their background have tended to demonstrate greater assessment literacy (Alkharusi, 2011b; Zhang & Burry-Stock, 1995, 2003), and teachers who scored well in an educational measurement course held more competence in assessment than their peers who earned lower grades (Alkharusi, 2011a). No clear consensus has been reached on the relationship between teaching experience and level of assessment literacy. In some cases, a positive relationship has been found (Alkharusi, 2011b; Hoover, 2009; Mertler, 2004; Zhang & Burry-Stock, 1995), in other cases that relationship has not revealed itself (Kershaw, 1993; King, 2010; Zhang & Burry-Stock, 2003). Teachers at the secondary level demonstrated more assessment literacy than teachers at the elementary level (Hoover, 2009; Quilter & Gallini, 2000). Teacher gender and college major (traditional core subjects vs. arts, life skills, and physical education) showed no relationship to assessment literacy, in a single study (Alkharusi, 2011b).

The relationships between psychological constructs and assessment literacy have also been investigated, again to mixed results. Attitudes toward educational measurement have shown a moderate positive relationship (Alkharusi, 2011b; Quilter & Gallini, 2000) and nearly no relationship (Kershaw, 1993) to assessment literacy. Though only examined in one study, confidence in educational measurement was found to be related to assessment (Alkharusi, 2011b).

Finally, studies have also examined how assessment literacy relates to classroom outcomes such as use of assessment and student work. Such work is important for justifying academic attention given to the topic of assessment literacy and striving for meaningful impacts within the educational sphere. Arter (2001) found that instilling foundational assessment

knowledge in teachers could make them better users of assessment. McMorris and Boothroyd (1993) found teachers with higher assessment literacy produced higher-quality tests. Perceived competence in assessment was found to be related to the extent of use of both performance-based and pencil-and-paper assessments (Kershaw, 1993). Better use of assessments may or may not translate to improved outcomes for students. Tentative relationships were observed between higher assessment literacy and improved student scores on achievement tests (Braney, 2011; Lukin et al., 2004) and classroom work (Koh, 2011), but a lack relationship with student standardized test scores has also been found (Mazzie, 2008). Again, drawing conclusions about assessment literacy's relationships with classroom outcomes, across the body of literature, is complicated by the fact that several findings may only be attributed to a single study.

Human dimensions of assessment literacy scholarship

Among 126 authors across all of the works, only 16 appeared on more than one study. Craig Mertler was the most prolific contributor, appearing on five studies. Hussain Alkharusi appeared on four studies. These two men most displayed attempts at sustaining a line of research related to assessment literacy. Other cases of authors producing multiple works were either too inconsistent (e.g., Zhang, 1996; Zhang & Burry-Stock, 1995, 2003) or arose out of a single context(e.g., University of Colorado Assessment Project, Borko, 1997; Borko, Davinroy, Bliem, & Cumbo, 2000; Borko et al., 1997).

Author affiliation was examined to see what kinds of institutions of higher learning contributed to the assessment literacy knowledge base. These institutions were categorized by the basic classification of the Carnegie Foundation for the Advancement of Teaching. The classification of the first author's institution was recorded. In cases where the first author was not affiliated with a university in the United States or with a university at all (e.g., a public schools

system) affiliation of the second author was recorded. Among the 50 studies available for classification, 27 were carried out by authors at universities labeled “very high research activity”, the Carnegie Foundation’s highest rating in terms of research orientation. Another 16 were carried out at “high research activity” institutions; 6 were conducted by authors at universities offering a Master’s degree as the highest degree available; 1 author was at a Baccalaureate-only institution.

Nearly a quarter (16 of 65) of the studies were conducted either with a sample outside of the USA or by researchers with foreign affiliations. The country with the highest representation in terms of first author affiliation was Canada, with 6 studies, followed by Oman (4), Taiwan (2), and Australia, Mexico, Singapore, and the United Kingdom (1 each). One of the studies by a Canadian author was completed with a sample of participants from Kenya (Chirchir, 1995). All other studies were conducted in the country of author institution. The representation of studies with an international affiliation is interesting given the assessment climate in the United States relative to other countries. The U.S. possesses a pervasive accountability-driven testing environment in which assessment results directly impact teacher evaluation and employment, yet 24% of English-language assessment literacy studies, most of them published or presented in the U.S., over the last 20 years were conducted in a context that does not carry stakes as high for teachers as one can find in the U.S.

Discussion

The purpose of this study was to provide a basis for establishing a collective memory in the area of empirical assessment literacy study and identify gaps and inefficiencies in attention. A systematic review of the literature found 65 published works to fulfill this aim. In general, there were some encouraging signs (e.g., attempts to measure assessment literacy with

previously vetted instruments) and signs that the study of assessment literacy is not reaching its full potential (e.g., lack of findings that were repeated and consistent across studies). In this section I will review this complexity within the literature, and identify key areas to address within the study of assessment literacy.

Measurement of assessment literacy

The use, extension, and adaptation, by multiple authors, of the instrument based on the work of Plake, Impara, and Fager demonstrates good practice. Doing so helps to build a consistent base of knowledge. The field should be careful, however, about relying solely on the TALQ and its offspring lest it succumb to a mono-operation bias (Shadish, Cook, & Campbell, 2002). For example, studies using the TALQ or Assessment Literacy Inventory found teachers struggled with communication of assessment results. An inspection of the questions used across these studies, however, reveals that only one question tests teachers on strategy for communicating a particular result. Other questions targeted toward this standard address it indirectly, by asking the teacher essentially to choose the correct definition of a term. The use of other *Standards*-based instruments is good because they can operate from a common test blueprint, but provide unique coverage of the construct.

While adhering to a set of published professional standards is a defensible way to operationalize the assessment literacy concept, there are drawbacks. For one, the acts of administering, scoring, and interpreting assessment results are all captured by a single standard. One could argue the underlying skillsets for each of these acts, however, are distinct, especially for interpretation vis à vis the other two. Using the Standards as a blueprint for instrument development may underemphasize important competencies.

A second drawback is that instruments based on the Standards target what experts say teachers *ought to* know. Researchers should also consider basing their measurements on the kinds of tasks teachers are routinely asked to carry out. An empirical investigation via job task analysis could inform development of such instruments. The field should also consider the untapped potential assessment literacy may hold in relation to the activities of teachers. Instrument development could consider not only the current tasks of teachers, but the tasks in which they may engage with a solid understanding of assessment. In this case instrument development could be informed by a combination of job task analyses, perhaps with teachers who are experts in assessment, and judgment of the professional community (e.g., Brookhart, 2011a; Popham, 2009).

Furthermore, there is justification for measuring assessment literacy beyond what can be captured on a test of assessment knowledge. A disconnect has been demonstrated between the declarative and procedural competencies of assessment literacy. Observations of practice have revealed teachers may not be able to demonstrate the assessment literacy they show on the TALQ (Scribner-Maclean, 1999). Understanding of the principles and purposes of assessment and the tools available for assessing may not be reflected in actual assessment practice (Siegel & Wissehr, 2011). Such gaps between declarative and procedural knowledge have been observed and accounted for by researchers studying literacy in areas such as public law (Engel, 2008) and the World Wide Web (Page & Uncles, 2004). Assessment literacy researchers should develop measurements that tap into knowledge not only of "what" but also "how". Thorough measurement within the domain should also consider common behavioral antecedents, such as self-efficacy (Bandura, 2001), attitudes and subjective norms (Fishbein & Ajzen, 2010), locus of control (Ng, Sorensen, & Eby, 2006; Rotter, 1966), and personality (McCrae & John, 1992;

Poropat, 2009), to name a few examples. Some of these constructs have indeed been measured within the assessment literacy realm (Betz, 2009; DeLuca & Klinger, 2010; Leighton, Gokiert, Cor, & Heffernan, 2010), however, such measurement has often been done in the absence of concurrent measurement of assessment knowledge. Doing both in tandem is necessary for building understanding about the links between such constructs and effectively improving assessment competencies in a meaningful way.

Lack of cohesion and direction

The issue of teachers working with test results, and their capacities to do so skillfully, has received attention through the years, but the attention has been spotty and lacking a consistent effort on all required fronts. What has been particularly lacking is empirical follow through on the calls of certain scholars who advocate for more attention on these teacher competencies. The present study identified 36 articles published in peer-reviewed journals over a 21-year period (i.e., less than 2 articles per year). Such an output is insufficient for sustaining a growing body of knowledge, as scholars must call on potentially out-dated literature to support their studies, and lack a diversified context (i.e., known findings across samples, types of interventions, etc.) in which to situate their investigations.

Calls for assessment literacy have been couched in such notions as teacher decision-making (Airasian & Jones, 1993), teacher evaluation (Popham, 2006), core job functions (W. D. Schafer, 1991), and "dire consequences for students" (Stiggins, 1999), implying some impact of assessment practices on student learning. Yet few studies have aimed to make that connection. The connection, understandably, is a difficult one to make as there are several logical arguments that must be strung together to tie teacher assessment literacy to improved student achievement. As a brief example, one such argument may posit that assessment literate teachers have the

capacity to understand what assessment results can tell them about the performance of their students, and will be able to communicate the results more effectively to their students, who will then have a clear understanding of their academic strengths and weaknesses and be able to more efficiently focus their efforts to learn, thus reaching higher levels of achievement. Another argument could say that the teachers who have the capacity to understand assessment results will know how those results can be applied toward optimum instructional strategies, which in turn will lead students to higher achievement.

As noted in the present review, studies have drawn a tentative correlation between assessment literacy and student achievement, but these studies did not possess designs powerful enough to rule out the many rival hypotheses that could be developed for why gains in student achievement were seen (e.g., novel stimuli, improved support structures in the schools, other in-service trainings completed by teachers). What is more troubling is that the scholarly community has not taken upon itself the duty to build understanding of assessment literacy in such a way that it can serve as the foundation for the logical arguments laid out above. Assessment literacy studies exist in largely in isolation, answering inwardly focused research questions, and the research community has done little to provide solid backing for any claims that assessment literacy directly impacts outcomes of the highest public priority.

Along these lines much of the research base is on shaky footing with regard to justifying its existence. Articles in this review (Alkharusi et al., 2011; Daniel & King, 1998; Quilter & Gallini, 2000) as well as others in the broader discussion of assessment literacy (e.g., W. D. Schafer, 1993), to provide rationale for their respective works, utilize the claim that teachers spend as much as a third of their professional time engaged in assessment activities. The citation given is Stiggins and Conklin's (1988) report, *Teacher Training in Assessment*, a product of the

Northwest Regional Educational Laboratory (NWREL). It is understandable why this claim would be popular, as it apparently uncovers an aspect of the teaching profession that is not as visible as classroom instruction and lesson planning, but integral to the successful performance of a teacher. Unfortunately, if one goes to the source, and thoroughly inspects the NWREL report, the claim of how much time teachers spend engaged in assessment-related activities is made without any backing. This is a major crux of many arguments, but fails to pass scrutiny. To advance the study of assessment literacy and have it able to face challenges of its worthiness, we need stronger evidence than appeals to authority.

The calls for assessment literacy remain strong, but scholars have demonstrated a lack of scientific follow through. More attention is warranted on measurement of assessment competence that addresses the full spectrum of qualities that reflect what it means to be literate. Content justifications for employed measures are sound, but we are limited in what we know about assessment literacy (e.g., dimensionality) by the scant psychometric work that has been published in the area. The field needs more commitment from those with an interest in the topic, and work needs to be done more systematically. With these issues addressed, we should see more advancement in the next 20 years than we have in the last 20 years.

CHAPTER THREE

FACTOR STRUCTURE OF A TEACHER EDUCATIONAL MEASUREMENT SELF-EFFICACY SCALE

In the current accountability climate of education in the United States, an effective educational system depends on professionals that are literate in assessment and can take the appropriate actions in response to test results. Standardized testing is pervasive in state and district assessment systems, and often carries high-stakes consequences. Students may be denied graduation, promotion, or access to beneficial learning resources/environments (Center on Education Policy, 2007; Darling-Hammond, 2007; Heubert & Hauser, 1999; Penfield, 2010). Teachers may have bonus pay withheld or lose their jobs all together (Greenhouse & Dillon, 2010; Sawchuk, 2011). School administrators may lose their jobs as they are replaced in "turnarounds", resulting from the awarding of School Improvement Grants (U.S. Department of Education, 2010, p. 12). It is imperative that all professionals in the educational system be able to understand test results to the extent that they can evaluate student and school performance, and plan a course of action to achieve or maintain optimal performance.

At issue, here, is an appreciable amount of literacy related to test scores and how those scores came to be. Such competency could be described as educational *measurement literacy*. A distinct subset of assessment literacy (Stiggins, 1991), measurement literacy is aimed specifically at the use and interpretation of standardized tests. Following the model established in reading literacy (Lonigan et al., 2008, p. 75), to be measurement literate, possession of a portfolio of skills and knowledge is necessary, but insufficient on its own. One also must consider a behavioral component, a kind of internalization of competencies and an orientation to make using assessment results a habit. The picture of literacy is completed by what the individual does

with acquired technical skills, how frequently those skills are called upon, and to what ends they are used (Cremin, 1988). Within the assessment system realm, the application of these skills comes in the form of test score use. The educational workforce historically has not received training in educational measurement at a level to match the impact measurement activities have within the contemporary educational climate (Roeder, 1972; W. D. Schafer & Lissitz, 1987; Stiggins, 1999; Wise, Lukin, & Roos, 1991). Therefore, more attention to the topic of the understanding and use of the outputs of testing systems in warranted. The purpose of this study is to examine the internal structure of a measure of teachers' self-beliefs about their abilities to carry out basic uses of test scores, to evaluate the instrument's ability to support inferences regarding this characteristic.

Social cognitive theory

If one accepts that literacy levels are not at a requisite level, then maximal use of test results has not been achieved. Test score use is a behavior. Therefore, to reach a maximal use of test results we ultimately must change a behavior, and must look to behavioral theories for guidance on what moves teachers to act or not to act in response to student assessment results. What are the antecedents, and what are the mechanisms through which behavior is determined? The answers to these questions will determine focus of efforts to improve test score use. The behavior of interest is one that is under volitional control. That is, within the confines of what a teacher is allowed or required to do by authority, the attention and effort one gives to using test scores would be largely determined by that individual. Therefore it is likely this behavior will result from conscious evaluation of various motivational factors. One such factor is how one expects to benefit from an action. Expectations of certain outcomes arising from behavioral choices and evaluative judgments of those outcomes can drive an individual's behavior

(Wigfield, Tonks, & Klauda, 2009). Complementing considerations of expectations are self-perceptions about the ability to arrive at a certain outcome. Self-perceptions about one's capabilities are incorporated under the theme of *self-efficacy* in social cognitive theory (Schunk & Pajares, 2009). Rather than evaluating one's ability to carry out a future task, self-efficacy beliefs target how one assesses current competence.

Application of social cognitive theory suggests feelings of unpreparedness could preclude teachers from acting upon the results they are presented. That is, low self-efficacy can lead to low organization and execution of actions (Pintrich & Schunk, 2002). It reasons that low efficacy beliefs related to interpretation and explanation of educational measurement concepts could dissuade teachers from actively engaging with standardized test results. Such inaction limits the effectiveness of summative and interim assessment systems. Thus, information is needed on teachers' self-efficacy in the area of measurement to begin to systematically influence the situation.

Self-efficacy measures that achieve maximum explanatory and predictive power are those matching the level of specificity of the domain of functioning being analyzed and reflecting the various task demands within that domain (Schunk & Pajares, 2009, p. 50).The role of assessment strategies has been considered within investigation of the nature of more global conceptions of teacher self-efficacy (e.g., Ross, Cousins, & Gadalla, 1996; Siwatu, 2007), but study of self-efficacy for outcomes that are expressly assessment oriented would shed the greatest light on test score use. A few examples of such study exist in the literature. Contrasting results have been obtained, related to the levels of assessment self-efficacy in the teaching workforce. Experienced, secondary-level teachers have demonstrated high levels of assessment self-efficacy (Chapman, 2008). Studies of pre-service teachers have found both high (Ogan-Bekiroglu, 2009)

and low (Volante & Fazio, 2007) levels of self-efficacy. Evidence of the construct's malleability was found in a study of the outcomes of a web-based professional development program for teachers (Huai, Braden, White, & Elliott, 2006).

Among these studies there was some lack of adherence to strict theoretical notions of self-efficacy and its measurement, as the scales dubbed *self-efficacy* took the form of self-ratings of assessment competence (e.g., Huai et al., 2006; Volante & Fazio, 2007), rather than self-beliefs related to organizing and carrying out a course of action to achieve a given outcome (Bandura, 2006). As a whole, these measures of assessment self-efficacy, much like measures of assessment knowledge (e.g., Mertler & Campbell, 2005; Plake et al., 1993), have been oriented toward classroom assessment activities. Self-efficacy for actions related to standardized testing outputs has not received the same attention. Furthermore, published psychometric analysis of assessment self-efficacy scales has largely been limited to basic descriptive statistics, (i.e., means and standard deviations), internal consistency reliability estimates, and basic analysis of contrasting groups. One study (Ogan-Bekiroglu, 2009) found evidence for a four-factor structure using exploratory factor analysis, but the description of the method for this analysis was very brief, and not enough information was provided to evaluate the appropriateness of the means used to determine the number of factors (Eigenvalues > 1, percent of total variance extracted) nor the extent to which these methods were in agreement. No further information about the factor analysis (e.g., rotation and extraction methods, factor pattern and structure coefficients) was included in the article. The present study aims to fill the noted shortcomings in the literature.

Method

Participants

Two random samples of teachers (N_{total} = 5000; $n_{sample1}$ = 3000; $n_{sample2}$ =2000) employed in Washington State elementary schools were selected to participate in a web-based survey. The selection of Washington was in keeping with the land grant mission of the university. The Tailored Design Method (Dillman, Smyth, & Christian, 2009) was employed for conducting the survey. This method utilizes social exchange concepts to encourage participation, and is characterized by its flexibility to adapt to the specific survey exercise. Communications sent via email to the participants included a pre-notice letter, invitation to participate, and two reminders. For both samples, the pre-notice letter was sent 2 days before the invitation to participate. For sample 1, reminders were sent at 6 and 13 days after the invitation. For sample 2, reminders were sent at 11 and 32 days after the invitation.

The overall response rate among sample 1 was 29%. An effective sample of 787 teachers was obtained after removing respondents who provided no responses to the self-efficacy items. Among this sample 93% of the respondents were Caucasian and 86% female, closely resembling the population proportions of 92% and 83%, respectively (State of Washington Superintendent of Public Instruction, 2008a, 2008b). Additionally, 73% of the respondents held a master's degree, and another 13% had completed some graduate school. Respondents averaged 15 years of teaching experience. Among elementary school teachers in the population, approximately 65% had earned a master's degree, and the approximate average years of teaching experience was 12.2 years (State of Washington Superintendent of Public Instruction, 2011). Therefore the teachers in sample 1 had slightly more education and experience than the population.

The overall response rate among sample 2 was 26%. An effective sample of 499 teachers was obtained after removing respondents who provided no responses to the study items. Among this sample 92% of the respondents were Caucasian and 84% female. A master's degree had been earned by 73% of the respondents, and another 17% had completed some graduate school. Respondents averaged 16 years of teaching experience. Similar to sample 1, the teachers in sample 2 were, on average, more highly educated and had been in the profession longer.

Instrument

To address the need for empirical studies of teacher self-efficacy as a component of educational measurement literacy, the Teacher Educational Measurement Literacy Scale (TEMLS) was developed with not only a test of understanding of concepts but also an assessment of how well teachers believe they can carry out various testing related tasks. This self-efficacy scale (TEMLS-SE) is the only known measure of measurement literacy within teacher self-efficacy circles. The scale consists of 21 items assessing a teacher's judgment of his or her capabilities to use test score information. Items were modeled after Bandura's (2006) guidelines for developing self-efficacy assessments. The items employed the stem, "How well do you believe you can..", and were rated on a 7-point scale (1 = Not at all well, 7 = Very well). Rating scale construction diverted from the recommended 11-point scale using 0 and 100 as endpoints because 7-point scales have been shown to maximize reliability and accuracy of responses (Lozano et al., 2008). For this study, the TEMLS-SE items were embedded in a questionnaire used for a larger study that asked teachers to complete a battery of items testing their understanding of measurement concepts, and report on their experience with measurement training, interests in further training, and basic demographic information. In total, the

questionnaire consisted of 50 items, and required an estimated average, based on pilot testing and respondent feedback, of 20-25 minutes to complete.

Analysis

This study implemented three components of analysis. The first component involved conducting an exploratory factor analysis on a randomly-chosen portion (n=394) of sample 1. The second component used the remaining portion of sample 1 (n=393) to conduct a confirmatory factor analysis of the best-fitting structure(s) revealed in the first analysis. The third component involved a cross validation of the factor structure obtained from the second analysis using the entirety of sample 2 (n=499).

Assumptions. Standard factor analytic procedures assume sufficient sample size, approximately interval-level scales, and multivariate normality (Brown, 2006, p. 107). The available sample size (n=394) for the exploratory factor analysis was sufficiently large to handle a situation of low communalities (i.e., .20-.40), and low numbers of indicators per factor (e.g., 3 factors with 3-4 indicators each; MacCallum, Widaman, Zhang, & Hong, 1999). Power analysis for testing the latent factor models assuming conservative cases (less than ideal fit) resulted in a range of N from 170 to 514 for power of 0.80 (Hancock, 2006). Therefore the samples available for initial confirmation of a factor model (n=393) and cross-validation of that model (n=499) were of sufficient size. The data available for analysis were at the item level, reflecting respondents' self-evaluations on a 7-point scale. Such a response scale has demonstrated qualities of interval-level measurement and frequently served as the basis for factor analyses (Floyd & Widaman, 1995, p. 288). Multivariate normality was assessed via inspection of skewness and kurtosis among the scores for each factor, and Mahalnobis distances and multivariate kurtosis. Data exhibited pronounced non-normality (e.g. multivariate kurtosis=10.3;

skewness and kurtosis >|1| for some items), so an alternative estimator, robust ML, and fit statistic, Satorra-Bentler χ^2 was used (Brown, 2006, p. 379).

Missing data. On average, across variables, there was a missing rate of 2.2%. Analysis of respondent data suggested that missing responses could be considered at least missing at random (MAR). Multiple imputation (MI) in the presence of MAR data has demonstrated the production of accurate and efficient parameter estimates (Allison, 2003; Schafer & Graham, 2002). Simulation studies have demonstrated that MI produces less model convergence failures (Enders & Bandalos, 2001; Newman, 2003; Olinsky, Chen, & Harlow, 2003), less bias in parameter estimates (Arbuckle, 1996; Enders & Bandalos, 2001; Gold & Bentler, 2000; Newman, 2003; Olinsky et al., 2003; Wothke, 2000) and better efficiency of parameter estimates (Enders & Bandalos, 2001; Wothke, 2000). As a collection, the work demonstrates that employing MI can result in more accurate factor analyses. This demonstration is especially important in the context of a factor analysis like the present study, in which such analyses are employed to gather validity evidence.

Exploratory factor analysis. Factor analysis identifies patterns of responses in a large dataset, by investigating multivariate intercorrelations among items (Nunnally & Bernstein, 1994). The process seeks to explain common variance among items via a factor model, thus reducing the data to a specification of relationships among items, factors, and presences of error. The inference is that common variance is driven by latent characteristics of the respondent (e.g., a single or multiple facets of self-efficacy). Therefore patterns of response are thought to reveal an underlying model of the construct(s) measured by a psychoeducational instrument. Exploratory factor analysis (EFA) is entirely data-driven, as no a priori model is set forth. It is suitable for initial analysis of the internal structure of the TEMLS-SE because item writing for

this scale followed a general consideration for common teacher actions, and did not adhere to any pre-specified categorization of these actions (Floyd & Widaman, 1995). At the outset, I was not testing any hypotheses about the structure of the data, and EFA provides the correct tool for this type of endeavor.

The EFA was conducted using the FACTOR procedure in SAS 9.2 (SAS Institute, Inc., 2008). Principal axis factoring was used for initial extraction of factors. These initial factors then underwent promax rotation to arrive at a factor solution. Multiple criteria were used to determine the number of factor to retain, as suggested by Fabrigar, Wegener, MacCallum, and Strahan (1999). Parallel analysis, inspection of Scree plots, and examination of the residual correlation matrix, guided the retention process. Interpretability and usefulness of the final factor solution was considered as well. Regarding individual item retention, ideally, all items will have a factor pattern loading of ≥ 0.30 on only one factor (i.e., the resulting factor model will possess simple structure; Comrey & Lee, 1992). Additionally, alpha value changes resulting from the deletion of an item should be negative. Items not displaying high degrees of association with their primary factors and not contributing positively to internal consistency estimates may be removed unless deemed essential to content coverage, and another iteration of the EFA will be conducted. Alternatively, the item(s) will be retained and their respective models will be compared in the confirmatory factor analysis framework.

Confirmatory factor analysis. Whereas EFA is data-driven, confirmatory factor analysis (CFA) relies on a priori specification of hypothetical models. The analytical framework then submits these models to statistical testing. CFA is appropriate when researchers believe an instrument measures a construct with a meaningful dimensional structure, and hope to inform theoretical discussions of the construct (Floyd & Widaman, 1995).

Robust maximum likelihood (robust ML) estimation via LISREL 8.8 (Jöreskog & Sörbom, 2006) was used to test internal structure of the covariance matrix of item responses. Model specification was determined by the outcomes of the EFA. The best models from the EFA were tested against a one-factor model to affirm that a multiple-factor model significantly improves data fit. Additionally, a two-level model was tested, where all first-order factors indicated a single second-order factor representing an overall measurement literacy competency. Such a model should be tested in the presence of correlated first-order factors (Thompson, 2004), a situation common among the self-efficacy concept (e.g., Bosscher & Smit, 1998; Chen, Greene, & Crick, 1998; Pajares, Hartley, & Valiante, 2001; Tschannen-Moran & Woolfolk Hoy, 2001). Furthermore, more exploration of second-order models has been advocated, specifically within the realm of teacher self-efficacy research (Henson, 2002).

Model evaluation. Several indicators were used to assess model fit. The chi-square test provided a measure of fit in terms of statistical significance ($\alpha = 0.05$). The Satorra-Bentler chi-square (SB χ^2) statistic was inspected in addition to the maximum likelihood chi-square (ML χ^2), as the SB χ^2 has been shown to behave well in conditions of nonnormality (Curran, West, & Finch, 1996). The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were used to indicate the difference in fit of the null and target models relative to the fit of the null model. The SRMR represented the standardized difference between the observed covariance and predicted covariance. Following the recommendations of Hu and Bentler (1999), CFI and TLI values of 0.95 or greater and SRMR values of less than 0.08 provided cutoff criteria for well-fitting models.

In addition to consideration of fit indices, models were examined for the presence of other indicators that could signify lack of fit (e.g., residuals, interfactor correlations, high

structure coefficients on non-specified factors). Well-fitting models should possess standardized residuals that follow a relatively normal frequency distribution and do not exceed an absolute value of 2.00 (Brown, 2006, p. 118). Interfactor correlations exceeding .85 indicate a lack of discriminant validity evidence (Brown, 2006, p. 166), and in such cases more parsimonious solutions (i.e., models with fewer factors) may be sought. Finally, inspection of factor pattern and structure coefficients should reveal that items load saliently on only one factor. Items deviating from this pattern were reviewed, and either considered for deletion or allowed to indicate more than one factor, depending upon conceptual meaningfulness. Maximum modification indices provided in the software output were considered both in terms of influence on model fit and conceptual meaningfulness.

Cross validation. To see if the confirmed factor structure would generalize across samples of teachers, the model with the most acceptable fit was cross-validated with the second half of the sample using an invariance testing framework. This framework involved a series of tests, using nested χ^2 methods, that examined equivalence of a) model structure, b) factor loadings, c) intercepts, and d) residual variances (Brown, 2006, pp. 269–270). A SB χ^2 difference test (Satorra & Bentler, 2001) was conducted using SBDIFF.EXE (Crawford & Henry, 2003) to assess incremental loss of model fit from introducing additional constraints across samples. Significant results from the test suggested unacceptable loss of fit. Additionally, decrement of other fit indices (e.g., CFI) was subjectively evaluated. Decreases in CFI and TLI of more than .01 and increases in SRMR and RMSEA of more than .01 were considered unacceptable for satisfying invariance requirements. Internal consistency reliability estimates were calculated for the scores on the factors for the final model across both groups. Given the opportunity for cross

validation, modification of the measurement model in the first phase of the confirmatory analysis was liberal, though still rooted in consideration of item content.

Results

Exploratory factor analysis

Results were mixed in terms of determining a factor solution for the TEMLS-SE items through exploratory processes. Strict simple structure (i.e., item factor pattern loadings of >= .30 on only one factor) could not be achieved, however it was generally the case that each item had one factor that it was clearly most associated with. Exceptions included items 11, 12, 13, and 14 in the 2- and 3-factor solutions. A 4-factor solution came the closest to achieving simple structure, including resolving the issues with the aforementioned items, but suffered from interpretability problems. Specifically, three items—2, 4 and 5— that loaded on the first factor in the 2- and 3-factor solutions were extracted into their own unique factor. While one could reasonably argue for some uniqueness among items 4 and 5, relative to the rest of the scale, I could not provide a meaningful interpretation for why item 2, *How well do you believe you can explain a scale score to a student's parent*, would associate with these items independent of item 3, *How well do you believe you can explain a percentile score to a student's parent*, for example, which also targets the explanation of types of test scores to parents.

To achieve simple structure, items were considered for deletion. The best candidate for such deletion was item 14, *How well do you believe you can interpret a student's strengths and weaknesses based on his/her mastery level*, because there was not a large difference between its pattern coefficients in the 2-factor model (0.38 on factor 1 and 0.45 on factor 2) and in the 3-factor solution it did not possess a factor pattern loading greater than 0.40 on any factor. Attempts were made to find a solution with item 14 removed from the analysis, but that left the

3-factor solution with one factor possessing only two items (15 and 16). Such a situation is not ideal, regarding concerns for factor identification (Brown, 2006, p. 72). Therefore, based on factor loadings, the 2- and 3-factor solutions appeared to be the strongest candidates for representing the data.

Scree plots indicated a strong primary factor, followed by between one and four additional factors, depending on the imputation iteration. Parallel analysis revealed Eigenvalues of the reduced correlation matrix did not exceed Eigenvalues produced by a correlation matrix of random data after two factors. The average off-diagonal residuals (i.e., results of subtracting the factor correlation matrix from the actual correlation matrix) were 0.076 for the 2-factor solution (90% above .05, 10% above .10), 0.061 for the 3-factor solution (71% above .05, 0% above .10), and 0.046 for the 4-factor solution (33% above .05, 0% above .10), indicating moderate to low misfit across the solutions (Pett, Lackey, & Sullivan, 2003, p. 123).

In consideration of all of the information discussed, here, I advanced two potential models to the confirmatory phase of analysis—a 2-factor solution and a 3-factor solution. In the 2-factor solution, the factors were labeled *score/test processing* and *basic responsibilities*. The score/test processing factor contained items gauging a teacher's self-efficacy for understanding and interpreting basic scores and information regarding a test's technical quality, either for her own benefit or to explain to a student's parent. The basic responsibilities factor contained items gauging a teacher's self-efficacy for administering and scoring tests and interpreting a student's assessment performance to judge proficiency or target instruction. These factors correlated at a .53 level for the sample. In the 3-factor solution, a *score transfer* factor was introduced, and contained items from both of the factors from the 2-factor solution. Score transfer items specifically targeted a teacher's self-efficacy for taking external test results and using them to

evaluate a student and target instruction. Interfactor correlations ranged from .50 to .55. Tables 3-1 and 3-2 outline the association of items with factors in each solution as well as resulting internal consistency estimates of reliability (i.e., coefficient alpha; Cortina, 1993; Cronbach, 1951). Internal consistency estimates were adequate across both factor solutions (i.e., >0.87; Cortina, 1993).

Confirmatory factor analysis

The results for all factor models tested are presented in Table 3-3. Based on comparisons of AIC values, both the two- and three-factor models provided better fit to the data than a one-factor model. We can therefore conclude that the TEMLS-SE data do not reflect a single dimension of self-efficacy. The three-factor model, SB χ^2 (186) = 1306, $p < 0.01$, CFI = 0.94, TLI = 0.94, SRMR = 0.14, appeared to fit the data better than the two-factor model, SB χ^2 (188) = 1521, $p < 0.01$, CFI = 0.93, TLI = 0.92, SRMR = 0.15. Because a single-order three-factor model and a higher-order model with all three first-level factors indicated by a single second-level factor would produce identical fit to the data, no higher-order models were tested. The three-factor model examined here remains at the first-order, with correlations among the factors specified. One could, however, specify a higher-order model if there was desire to use not only the three factor scores, but also a single composite score.

While the three-factor model provided better relative fit, it still fell short of all the pre-specified fit criteria. Therefore, modification indices were considered to try to improve the fit of the model to the data. The first modification was to allow item 11, *How well do you believe you can identify whether or not a student's test score meets a specified standard*, to cross load on the score/test processing and basic responsibilities factors, as there was a modification index for this cross-loading that was much higher than for any other lambda-x values. Other modification

indices for error terms were higher, but the centrality of the item-factor specifications to the model was given priority. Item 11 had shown factor pattern loadings of >0.37 on both of these factors in the exploratory analyses in both the two- and three-factor solutions. Additionally, a review of the item's content suggested the item did possess elements implied by both factors. With this justification, I allowed the item to cross load. Inspection of this new model revealed an interesting occurrence. When allowed to be indicated by both the score/test processing and basic responsibilities factors, item 11 actually showed a much strong affiliation with the former (factor pattern coefficient = 0.68) than the latter (0.16). This result was somewhat contrary to the results of the EFA, but again, the item did show a relationship with the score/test processing factor in that phase of the analysis. A subsequent model was specified, in which item 11 was indicated only by the score/test processing factor. Fit for this model SB χ^2 (186) = 1164, $p < 0.01$, CFI = 0.95, TLI = 0.94, SRMR = 0.13 was better in terms of the AIC, CFI, and SRMR. The TLI remained the same as the original 3-factor model. Only the CFI met pre-specified fit criteria, and so further modifications were considered.

Error terms were freed to correlate for three pairs of items. These modifications were considered iteratively, but, for brevity, are presented in a single step, here. The pairs of items—5 and 6, 9 and 10, and 19 and 20—each displayed content similarities to justify estimating correlations between their error terms. Items with similar wording can impose measurement effects (i.e., shared variance due to an outside cause) that necessitate the specification of correlated error terms in the model (Brown, 2006, pp. 181–182). Items 5 and 6 were the only two items that dealt specifically with the issue of understanding information presented in a test manual. Items 9 and 10 both involved comparisons of two scores from the same student. Finally, items 19 and 20 were the only items concerning administration of a standardized test. With these

pairs of item error terms correlated, SRMR remained high, and the SB χ^2, while again greatly reducing, still suggested a rejection of the model, SB χ^2 (183) = 703, $p < 0.01$, CFI = 0.97, TLI = 0.97, SRMR = 0.12, so the modification process continued.

After freeing the correlations of the error terms described above, high modification indices persisted for loadings of item 14 on both the score/test processing and basic responsibilities factors. Item 14 concerns the interpretation of a student's strengths and weaknesses based on his/her reported mastery level. Within this hypothetical task, the teacher would be required to carry out a basic responsibility of processing a type of score (e.g., scale score) and transfer interpretation of that score to inform a decision about a student (e.g., pass/fail). Therefore the item content does call upon elements of all three proposed factors. In the exploratory phase of analysis, factor pattern coefficients for item 14 ranged from .20 on the score/test processes factor to .36 on the score transfer factor. Given this pattern of results, two options were considered—eliminating the item from analysis and allowing it to load on all three factors. Model fit for retaining the item and allowing it to load on all three factors, SB χ^2 (181) = 654, $p < 0.01$, CFI = 0.98, TLI = 0.97, SRMR = 0.11, was clearly better than for eliminating the item, SB χ^2 (164) = 929, $p < 0.01$, CFI = 0.96, TLI = 0.95, SRMR = 0.12. This finding was desirable as the item's content was considered to address a fundamental task within measurement literacy.

While indices for further modifications remained high, such modifications could not be justified. It is possible that further modification of the model could have addressed the relatively large residuals that remained (SRMR=.11), but changes would not have been made on a theoretical basis and could have acted on idiosyncracies of the particular sample of data. So the model submitted for cross validation was the one depicted in Figure 1. This model met pre-

specified criteria for CFI and TLI, but SRMR and SB χ^2 were still high. The distribution of standardized residuals approximated normality, but with several residuals in the range of +/-2 to 4, and some extreme outliers (e.g., -10.2, 39.5). Interfactor correlations ranged from 0.38 to 0.62 (Table 3-4). Table 3-5 presents factor pattern and structure coefficients for the final model. Excluding those associated with item 14, factor pattern coefficients were at acceptable levels. Structure coefficients did signal some overlap between the factors. For example, item 8, How well do you believe you can interpret a student's score in relation to another student's score, held a structure coefficient of 0.56 with the basic responsibilities factor, which was higher than or equivalent to the pattern coefficients for four of the factor's six indicators. In another example, items 17 and 18 held structure coefficients with the score/test processing factor higher than three of the items specified to load on that factor.

Cross-validation of the factor model

The first step taken in cross-validation was to fit the derived model to the data from sample 2. Fit with the new sample was comparable to what was found with the initial sample, SB χ^2 (181) = 726, $p < 0.01$, CFI = 0.98, TLI = 0.98, SRMR = 0.12. The next step taken was to estimate fit for a multi-sample configural model (Byrne, 2012, p. 196) where the same pattern of freed and fixed parameters (i.e., model structure) was specified and estimated simultaneously for samples 1 and 2. This model provided a baseline of fit against which to compare more constrained models. The baseline model held fit properties similar to those observed with the samples individually, SB χ^2 (362) = 1392, $p < 0.01$, CFI = 0.98, TLI = 0.97, SRMR = 0.12, thus we have evidence that model structure is invariant across the samples. The first constraint placed on the model in the cross-validation process was to hold the factor loading values constant for both samples. This model displayed CFI and TLI statistics consistent with the baseline model,

SB χ^2 (385) = 1879, $p < 0.01$, CFI = 0.97, TLI = 0.97, SRMR = 0.25, but held a high SRMR and rendered a significant ($p<0.01$) SB χ^2 difference test. Tests of invariance among individual factor loadings revealed significant loss of fit ($p<0.01$) for about half of the factor loadings when they were held invariant one-at-a-time. No pattern could be discerned regarding the content of which items appeared invariant and which ones did not. Given a lack of invariance at the level of constraining factor loadings, no further steps were taken to cross-validate the derived model.

Discussion

The purpose of this study is to examine the internal structure of a measure of educational measurement self-efficacy, to evaluate the instrument's ability to support inferences regarding this characteristic in teachers and to explore the nature of the measurement self-efficacy construct, one that has not seen extensive attention in the literature. Results were inconsistent for the first aim, thus diminishing our ability to draw conclusions about the latter. Examination of model fit for TEMLS-SE data found good CFI and TLI results, but poor SRMR results. The SRMR findings reflect moderate residuals found in the 2- and 3-factor solutions in the EFA phase of analysis. Internal consistency reliabilities are high for each factor (i.e., >0.87; Cortina, 1993). The model, however, is hampered by localized misfit, such as a single item that loads unsatisfactorily on each of the three latent factors and other instances of low associations between item and factor (e.g., 0.49 for item 20 on the basic responsibilities factor).

The model demonstrated invariance in terms of structure, which is good considering the extent of modifications made to the measurement model. But claims of invariance could not be supported in terms of factor pattern values. Because no association between item content and invariance could be established, we are left to wonder why this level of invariance could not be established. While both sample 1 and sample 2 were drawn from public elementary schools in

the State of Washington, the sampling frame for sample 2 was more extensive. For sample 1, teacher contact information from 385 schools was available, whereas 550 schools, including the original 385, were available for sample 2. The growth of the teacher population available for sampling was more or less random (i.e., more schools districts with names starting with the letters A-P available), but it is possible that opening up the sample to be populated by a wider representation of the teaching workforce in the state introduced variability in teacher responses that could not be accounted for by the derived factor model estimates. Given that such minor fluctuations in the sample could inhibit invariance of the model, further study should attempt to fit and validate a factor model with a much more diverse (e.g., national) sample.

Despite inconsistent results and diminished opportunity for drawing conclusions, we nevertheless have a foundation upon which to build future inquiry. We are left with a window into the concept of measurement self-efficacy among elementary-level teachers. We have a model that displays pretty good fit and can be grounded in evaluation of TEMLS-SE item content. Overall, the results suggest the TEMLS-SE is capturing three distinct aspects of self-efficacy, therefore supporting claims made about teacher self-efficacy for measurement tasks and concepts within each of these aspects, but those claims could be supported to an even greater extent with some revision of items to attempt to achieve a cleaner, invariant solution. Such revision should be guided by further consideration of the measurement self-efficacy concept and the responsibilities teachers carry for working with student test results. These considerations could illuminate areas not substantially addressed by the TEMLS-SE items, and highlight potential ambiguities or points of confusion that may exist among the items. To further address these points, one could submit the TEMLS-SE items to a cognitive interview framework (Willis, 2005). Such an analytic technique has proven useful for understanding how respondents interpret

and complete instruments measuring concepts in the domain of motivation and behavior (e.g., Darker & French, 2009), and should therefore be suitable for implementation with a self-efficacy instrument. Item revision could improve how well the TEMLS-SE represents the underlying constructs, therefore improving fit of a measurement model to the data. Ultimately what is most important is to continue scholarly investigation of the matter of teachers working with the products of an educational system that is immersed in prodigious amounts of test data that directly impact so many.

Table 3-1.

Factors, their corresponding items, and internal consistency reliabilities from the 2-factor solution

Factor name	Corresponding items "How well do you believe you can…?"
Score/test processing α=0.93	i1. explain the median of a score distribution to a student's parent i2. explain a scale score to a student's parent i3. explain a percentile score to a student's parent i4. explain a grade equivalent to a student's parent i5. understand validity information presented in a test manual i6. understand reliability information presented in a test manual i7. interpret a student's score in relation to a state average i8. interpret a student's score in relation to another student's score i9. interpret a student's score on one test in relation to that same student's score on another test i10. interpret a student's score from one grade to the next grade i12. explain the content of a test to a student's parent i13. explain the purpose of formative assessment to a student's parent i15. interpret state test score report information to inform your teaching practice i16. interpret state test score report information to make a decision about a student
Basic responsibilities α=0.89	i11. identify whether or not a student's test score meets a specified standard i14. interpret a student's strengths and weaknesses based on his/her mastery level i17. interpret frequent in-class assessments results to inform your teaching practice i18. interpret frequent in-class assessments results to make a decision about a student i19. administer a standardized test to an individual student i20. administer a standardized test to a classroom of students i21. follow instructions on how to score a student's responses to a standardized test

Table 3-2.

Factors, their corresponding items, and internal consistency reliabilities from the 3-factor solution

Factor name	Corresponding items "How well do you believe you can…?"
Score/test processing α=0.92	i1. explain the median of a score distribution to a student's parent i2. explain a scale score to a student's parent i3. explain a percentile score to a student's parent i4. explain a grade equivalent to a student's parent i5. understand validity information presented in a test manual i6. understand reliability information presented in a test manual i7. interpret a student's score in relation to a state average i8. interpret a student's score in relation to another student's score i9. interpret a student's score on one test in relation to that same student's score on another test i10. interpret a student's score from one grade to the next grade i12. explain the content of a test to a student's parent i13. explain the purpose of formative assessment to a student's parent
Basic responsibilities α=0.89	i11. identify whether or not a student's test score meets a specified standard i17. interpret frequent in-class assessments results to inform your teaching practice i18. interpret frequent in-class assessments results to make a decision about a student i19. administer a standardized test to an individual student i20. administer a standardized test to a classroom of students i21. follow instructions on how to score a student's responses to a standardized test
Score transfer α=0.87	i14. interpret a student's strengths and weaknesses based on his/her mastery level i15. interpret state test score report information to inform your teaching practice i16. interpret state test score report information to make a decision about a student

Table 3-3.

Fit results for tested factor models

Model	Modifications	df	χ^2	SB χ^2	CFI	TLI	SRMR	AIC
2 factor	none	188	3459	1521	0.93	0.92	0.15	1607
3 factor	none	186	3100	1306	0.94	0.94	0.14	1396
3 factor	i11 on score/test processing	186	2782	1164	0.95	0.94	0.13	1241
3 factor	i11 on score/test processing; correlated error terms	183	1664	703	0.97	0.97	0.12	799
3 factor	i11 on score/test processing; correlated error terms; item 14 removed	164	2407	929	0.96	0.95	0.12	1021
3 factor (Sample 1)	i11 on score/test processing; correlated error terms; item 14 on all three factors	181	1545	654	0.98	0.97	0.11	754
3 factor (Sample 2)	i11 on score/test processing; correlated error terms; item 14 on all three factors	181	2147	726	0.98	0.98	0.12	826

Note: All χ^2 and SB χ^2 significant at α=0.01.

Table 3-4.

Interfactor correlations of the final 3-factor model

	Score/test processing	Basic responsibilities	Score transfer
Sample 1			
Score/test processing	1.00		
Basic responsibilities	0.62	1.00	
Score transfer	0.57	0.38	1.00
Sample 2			
Score/test processing	1.00		
Basic responsibilities	0.68	1.00	
Score transfer	0.63	0.43	1.00

Table 3-5.

Pattern and structure coefficients

Item	Score/test processing	Basic responsibilities	Score transfer
i8	**0.91**	0.56	0.52
i7	**0.84**	0.52	0.48
i10	**0.80**	0.50	0.46
i11	**0.80**	0.50	0.46
i9	**0.79**	0.49	0.45
i12	**0.74**	0.46	0.42
i4	**0.73**	0.45	0.42
i3	**0.70**	0.43	0.40
i13	**0.70**	0.43	0.40
i5	**0.64**	0.40	0.36
i2	**0.62**	0.38	0.35
i6	**0.61**	0.38	0.35
i1	**0.58**	0.36	0.33
i18	0.61	**0.99**	0.23
i17	0.60	**0.97**	0.23
i19	0.35	**0.56**	0.13
i21	0.32	**0.51**	0.12
i20	0.30	**0.49**	0.12
i15	0.60	0.00	**0.97**
i16	0.58	0.00	**0.93**
i14	**0.22**	**0.34**	**0.31**

CHAPTER FOUR

RESPONSE PROCESS EVIDENCE FOR THE TEACHER EDUCATIONAL MEASUREMENT LITERACY SCALE VIA THINK-ALOUD INTERVIEWS

The *Standards for Teacher Competence in Educational Assessment* (American Federation of Teachers et al., 1990) state that teachers, among other competencies, (a) should have the skills to administer, score, and interpret externally-produced assessments; (b) be able to use assessment results to plan their instruction and make decisions about individual students; and (c) be able to communicate assessment results to students, parents, and other lay audiences. Measurement concepts commonly employed on externally-produced assessments (e.g., scale scores, grade equivalents, standard error of measurement) appear to be often misunderstood by teachers and educational policy makers (Hambleton & Slater, 1997; Impara, Divine, Bruce, Liverman, & Gay, 1991), but a robust line of research on this topic does not exist.

In the general area of study of communication around test performance, attention has been given to how assessment results are communicated (e.g., Goodman & Hambleton, 2004; Jaeger, 1998; Wainer, Hambleton, & Meara, 1999), and effective ways of reporting test scores have been contemplated (Brown, 2001; Roduta Roberts & Gierl, 2010; Zapata-Rivera, Underwood, & Bauer, 2005). However, less attention has been given to individuals who actually receive the communications of assessment results, such as teachers, students, parents, and school administrators. The study of assessment literacy could be improved by providing an understanding about which educational measurement concepts—particularly those related to standardized, summative assessments—are understood well and understood poorly among today's K-12 teaching workforce.

To fulfill this aim we need to know the extent to which teachers can make appropriate interpretations of student performance, explain in plain terms that performance, and internalize the basic purposes of assessment in a way that makes it possible to integrate assessment results into instructional strategy. Scales to assess teachers' competencies in this area have been developed (e.g., Teacher Assessment Literacy Questionnaire, Plake et al., 1993; Classroom Assessment Literacy Inventory, Mertler, 2003, 2004; Mertler & Campbell, 2005; Assessment Knowledge Test, Wang, Wang, & Huang, 2008) with a broad orientation toward assessment, emphasizing teacher-developed classroom assessments. Strikingly, teacher knowledge of educational measurement concepts related to standardized testing, and particularly the measurement of such knowledge, has been given little attention in peer-reviewed studies since the NCLB initiative. A shift in attention is warranted to gauge teachers' knowledge about and understanding of measurement concepts found on score reports, and to allow such empirically defined information to drive the development of effective communication of assessment results and professional development materials.

To this end, the Teacher Educational Measurement Literacy Scales (TEMLS) were developed. These brief assessments—one for understanding of measurement concepts, one for measurement self-efficacy—may be capable of providing information about basic educational measurement literacy levels from a broad teaching population. The TEMLS have been piloted with a sample of Washington state elementary school teachers (Gotch & French, in press). Evidence has been collected from this population to evaluate the internal structure and score reliabilities of the knowledge assessment (Gotch & French, 2011). This initial work has been informative for modifications of the scale. In order for the TEMLS to function as an effective tool for investigating measurement literacy, however, a plethora of validity evidence is needed to

support the use of scores from the scale. This study gathers response process data to evaluate the extent to which the TEMLS knowledge scale can support valid inferences of teacher understanding of measurement concepts.

Response process investigation as validity evidence

The prevailing view of validity in the measurement community is that it pertains to the inferences drawn from the results of a test (Kane, 2006). Valid inferences can be supported by a chain of evidence that leads from the raw data to the inferences being made. Guidelines for what counts as evidence have been advanced by prominent scholars (e.g., Messick, 1989) and codified into the standards of the field (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) in the form of five main sources. One such source is the investigation of examinee response processes. Emerging from advancements in cognitive psychology, response processes are the only form of validity evidence that allows for analysis of the act of task performance, including direct observation of how an examinee interprets a given task, rather than relying on analysis of inputs or outputs of the task (Messick, 1989).

One method for discovering examinee response processes is the use of think-aloud protocols. In its most basic form, collecting verbal report data via think-aloud protocol involves having a research participant vocalize her thoughts while completing a psychometric instrument, in order to gain insight to how the participant understood each question and formed a response (Ericsson & Simon, 1993). This investigative process is important to evaluating the validity of claims made from the instrument because we demand some level of proof that the observable behavior of completing a test or survey item reflects the skills or characteristics the item is intended to measure. Conversely, using data from an instrument that has not been submitted to a

response processes investigation places a level of trust in the data that may or may not be founded.

The inclusion of response processes in the validity framework is rooted in the observation-interpretation-cognition triangle of assessment (Figure 4-1). A holistic explanation of the triangle posits that effective assessment and sound inferences depend upon meaningful connections between how we believe individuals build competence in a certain area, what we observe in terms of responses on an assessment, and how we interpret these responses in the context of the targeted domain (National Research Council, 2001, pp. 44–51). Understanding examinee response processes can reveal to us the extent to which the vertices of this triangle work consistently. In other words, to what extent are the interpretations we make from observing examinee responses reflective of the cognitive functioning we expect to be required by the test items? If all three are working together, we have stronger grounds for making valid claims from test data.

Some have cast doubts about the ability to tap into internal processes at work during task performance (Nisbett & Wilson, 1977), however substantial review of the literature suggests that verbal reports derived from think-aloud protocols can provide the evidence of interest when they follow a specific methodology and are interpreted within information-processing framework (Ericsson & Simon, 1993). For example, think-alouds have been able to capture, in real-time, detailed information about likes and dislikes, navigational processes, and interplay between these factors and one's use of resources in an e-learning environment, revealing issues that would likely not have turned up in the use of questionnaires or interviews (Cotton & Gresty, 2006). Protocol analysis of a pilot administration of an Assessment of Professional Behaviors by the National Board of Medical Examiners revealed inconsistent item interpretations and that medical

professionals would rate doctoral residents in absence of necessary information, based on general impressions or indirect evidence (Mazor, Canavan, Farrell, Margolis, & Clauser, 2008). Think-aloud procedures have also demonstrated that test developers must supplement expert judgment with evidence of examinee thinking processes, as the two sources of information do not always align in terms of identifying differences in item functioning across demographic groups (Ercikan et al., 2010).

A foundational component of the work proposed here is to understand, from the teachers' voices, what cognitive processes they move through as they respond to the TEMLS questions. Investigation of response processes is the only form of evidence where the voices of the teachers directly communicate to the researcher the meaning and interpretation of assessment items. Despite this fundamental quality, response process validity evidence is rarely collected on assessment instruments (e.g., <1.8% of published test manuals; Cizek, Rosenberg, & Koons, 2008). By focusing on response process evidence, the present study follows through on the validity aims established by the field, and provides a practical example of translating conceptual validity discussion into concrete methods, data, and conclusions.

The purpose of this study is to advance a line of research focused on teacher educational measurement literacy by investigating the mental processes teachers engage in when responding to items on scales intended to assess this form of literacy (i.e., validity evidence via response process data). Proximal goals include refining these scales and gaining insight into common misconceptions within the educational measurement domain. Distal goals include providing flexible professional development training for teachers to enhance skills in the said areas, especially as these areas relate to test score reporting.

Method

Participants

A snowball sampling method (Miles & Huberman, 1994, p. 28) was carried out with initial contacts going either to school principals or individual teachers. In any case where an interview was conducted on school grounds during a school day, the building principal was notified ahead of time. Invitations to teachers explained the general purpose of the study, basic information about the interview, and a notice that the teachers would receive $50 remuneration for approximately an hour of their time (not including transit to the interview site).

The sample of participants included 13 elementary-level teachers from two sites in the state of Washington—Pullman and the Tri-Cities (specifically Kennewick and Richland). The focus was on elementary-level teachers because they are the ones engaged in regular conferences with students' parents and therefore possess the greatest need for an ability to understand and explain test results to this group of stakeholders. Furthermore, the TEMLS had been piloted previously with a sample of elementary-level teachers. Thus, continuing with this group provided for an accumulation of validity evidence rather than obtaining evidence with a new population.

Individual years of teaching experience and grade level taught were used to guide the sampling and recruitment. This design accounted for key variables by which teachers can differ related to the construct of interest, and ensures representativeness of the sample. I was interested in talking with new teachers (i.e., less than 7 years of experience), mid-career teachers (i.e., 7 to 14 years of experience), and experienced teachers (i.e., greater than 14 years of experience). The breakdown of years of teaching experience coincided with when major shifts occurred in relation to standardized testing. Statewide testing in Washington began in 1997 via the Washington Assessment of Student Learning (WASL), hence the interest in recruiting participants with

greater than 14 years of teaching experience (i.e., teachers who entered the profession before the state began its largest standardized testing effort). Sampling purposefully below and above the 7-year breakpoint ensured both teachers new to the profession and those with a moderate amount of experience will be included in the study. Seven years of experience also approximated when the State extended accountability testing to 3rd-grade, in response to No Child Left Behind mandates.

I was interested in talking with teachers responsible for grades above and below 3rd grade. The breakdown of grade level taught, other than providing an even distribution across the elementary grades, coincides with important transitions related to standardized testing. Prior to 3rd grade, standardized testing is not used by the state for accountability purposes. Standardized testing at the lower grade levels would be carried out primarily for the purposes of assessing children on their development of basic skills (academic and non-academic) to make decisions about student placement and targeting instruction. Figure 4-2 displays a plot of years of experience by current grade level taught for the 13 teachers in the sample. It should be noted that it was common for teachers to have been responsible for a few different grade levels during the course of their teaching career.

Teachers in the sample came from 8 schools across the two locations—3 in Pullman, 5 in the Tri-Cities. Any single school in those areas was represented no more than 3 times across the sample. In general, the schools experienced relative success in terms of their students' achievement on state accountability tests. In the previous academic year, median proportions of proficient students among the schools ranged from 63-83% across the 3rd-, 4th, and 5th-grade Reading and Math MSP tests (State of Washington Superintendent of Public Instruction, 2011). Every represented school but one had at least a 50% proficiency proportion on these tests. The

lone exception had between 33% and 47% of students meeting proficiency on the MSP tests. There were 12 female teachers and 1 male teacher in the sample. All teachers held primary responsibility for a mainstream classroom. One teacher taught in a bilingual program (Spanish and English).

Instrument

The TEMLS were developed to match the content of student score reports from several U.S. states, Canadian provinces, and commercial vendors, supplemented with items that addressed fundamental educational measurement concepts such as validity, reliability, and the aims of different kinds of assessments (e.g., formative, summative, norm-referenced, criterion-referenced). Concepts addressed by the TEMLS are well-represented in measurement and assessment texts that would be used in a teacher training program (e.g., Linn & Miller, 2005; McMillan, 2007; Popham, 2005; Woolfolk, 1995). Developed items were reviewed by an expert item writer and a panel of local school district personnel working on assessment and score reporting issues. These reviewers, collectively, provided technical expertise as well as familiarity with the target population for the TEMLS. The TEMLS contains both a knowledge scale and self-efficacy scale. As response process analysis of the two scales would take two distinct orientations, the present study concerns only the knowledge scale, for the sake of clarity.

The TEMLS knowledge scale consists of 20 multiple choice (4 options) items covering issues such as interpretation of standardized scores; scores in relation to one another within a student, across students, and across schools; and proficiency level interrelation. Example question stems included: (a) *Evan, a third-grader, obtained a percentile rank of 90 on a standardized reading assessment. This indicates Evan…*, (b) *Proficiency exams are primarily used for determining if…*, and (c) *If Mrs. B. wanted to know Elise's strengths and weaknesses on*

certain reading skills what type of assessment would be most helpful? In a pilot administration to elementary-level teachers in the State of Washington, data from the knowledge scale possessed internal consistency reliability, as measured by Cronbach's alpha, of 0.47 (Gotch & French, in press). Data structure was ambiguous, with potential support for a one- or two-factor model (cf. Gotch & French, in press, 2011).

Ericsson and Simon (1993) state that moderately difficult tasks are best for think-aloud exercises because of the cognitive demands they impose. Tasks should be difficult enough that participants have to reason through and answer, but easy enough that participants can still devote cognitive processes to answering and vocalizing. In the pilot administration of the TEMLS knowledge scale, teachers were able to answer, on average, 70% of the questions correctly. Furthermore, individual item percent-correct figures ranged from 52% at the 25^{th} percentile to 91% at the 75^{th} percentile. Therefore, the difficulty of the TEMLS is amenable to think-aloud protocols, and should facilitate the success of participants in meeting the challenge imposed by the data collection method, thus providing rich data for analysis.

Procedure

Before going into the field, the think-aloud protocol was piloted with an individual who was a licensed teacher and recent Ph.D. recipient from a college of education. The pilot session involved implementation of the complete interview protocol. Two members of the research team were present, one to run the protocol and one to record notes and provide feedback. These members debriefed after the practice round, determining it to be a success in terms of how comfortable I felt as interviewer and the protocol's ability to collect data at a level amenable to analysis. So then the step of contacting active, in-service teachers commenced.

In the interest of keeping the think-aloud task to a manageable length without having the participants become too fatigued, I aimed for 1-hour interviews. Initially, the plan was to split the TEMLS into two forms, and administer only half of the instrument to any given teacher. After the practice interview, however, I decided an hour was enough time to accommodate all the items without extensive fatigue. The knowledge items were still split into two sections, and appeared as the first and third sections of TEMLS items. The TEMLS self-efficacy items, not analyzed in this paper, comprised the second section. The arrangement provided variety and novel stimuli for the participants, which may have kept them interested and motivated through the length of the think-aloud task.

The think-aloud technique minimizes interviewer imposed bias and requires minimal interviewer training (Willis, 2005, p. 53). This technique, therefore, was feasible and had the potential to provide information that will contribute to validation of the TEMLS measurement knowledge items. A challenge of the think-aloud technique is that it places a lot of burden on the research participant in terms of being able to recognize, vocalize, and articulate mental processes. To address this challenge, I tried to cultivate a friendly, comfortable setting, removing social norm concerns by placing the emphasis on the instrument (not on the participant or interviewer as author of the instrument), and providing a rehearsal opportunity before having the participant review the TEMLS will be implemented to address this challenge. Interview spaces were as free from distractions as possible, so that the teachers could focus their mental energies on the verbal reporting task. Teachers were informed that an audio recording of each interview would be made so that I could focus on the interview and not on note-taking. Each interview session included the teacher and me. In two interviews, another member of the research team was present to observe unobtrusively.

As stated above, the think-aloud procedure involves having research participants vocalize their thoughts as they move through a series of tasks. In the present study, those tasks took the form of TEMLS items. Each think-aloud interview consisted of three phases—warmup, concurrent, and retrospective. (See Appendix B for the interview protocol.) The warmup phase began after greeting the teacher, settling in to the interview space, and orienting him/her to the interview—its general purpose and structure. The teacher was then asked to speak for a couple minutes about how she got into the field of teaching, how long she had been at it, and what her experiences were. The purpose for this question was twofold. First, responses allowed for the collection of demographic information determined to be important to the sampling frame (i.e., years of experience and grade level taught). Second, by allowing the teacher to speak freely at the beginning of the interview, a rapport could be built between the teacher and myself, and a flow of conversation was established, which hopefully would lead to good verbalization when presented with TEMLS items.

Still within the warmup phase, the participant was presented three practice items. The first was an open-ended question that required application of knowledge of mathematical fractions to circle a set of shapes. The second item made use of a 7-point rating scale to familiarize the teacher with the form of the TEMLS self-efficacy items. The final warmup question was in multiple choice format, as the TEMLS knowledge items appear. It asked the teachers about the purpose of research, a topic area not directly addressed by the TEMLS, but close enough in relation that this warmup question was analyzed to provide confirmation of the verbal response patterns observed with the TEMLS items.

The concurrent phase involved the portion of the interview in which the participant was actively engaged with solving TEMLS items. In this phase, the thoughts of the participant as

they first entered consciousness were desired. It was essential to keep the participants vocalizations focused on short term memory actions during the concurrent reporting phase (Ericsson & Simon, 1993). I provided minimal prompts of a neutral nature (e.g., "keep talking", "what are you thinking about as you look at that question?") when the teacher lost the momentum of constant verbalization. (Prompts were provided after the teacher had been silent for several seconds.) Standardization of the process (instructions, prompts, etc.) was essential in order to have consistent impact across participants during the act of verbalization (Ericsson & Simon).

Retrospective reporting typically involves having the participant explain her problem solving strategies after completing a problem. A dual, concurrent/retrospective approach to the collection of verbal report data has been shown to provide the researcher opportunities to reduce the level of inference required of the researcher and improve completeness of the verbal report data (Taylor & Dionne, 2000). In the present study, however, it did not make sense to have teachers immediately explain how they arrived at their answers. For one, the number of items was numerous. Secondly, the items themselves were not very complex. All were in multiple-choice format, and most tapped into the teacher's knowledge of measurement terminology. A few questions required further application of this knowledge to a hypothetical scenario, but overall complexity of each item was minimal in terms of the strategies required to solve it. Finally, some teachers' verbalizations tended to move toward explanation after arriving at an answer rather than voicing these rationalizations while considering potential answer choices. In that sense, the teachers were already moving into a retrospective type of reporting.

So rather than ask the teachers to re-explain their process for arriving at an answer, following the think-aloud exercise, teachers were engaged in a conversation about the TEMLS

items and measurement and assessment topics and issues. While a full analysis of these conversations is beyond the scope of this paper, elements of these conversations will be highlighted where they provided evidence to confirm the findings of the present study.

Data analysis and results

As this is a validity study, the main goal was to know to what extent the TEMLS accurately captures a teacher's understanding of educational measurement as covered by the instrument's items. To that end we sought to answer two questions:

1) What are the mental processes that teachers enact when responding to items on the TEMLS knowledge scale?

2) To what extent do the response processes support validation of TEMLS knowledge scale scores?

Mental processes. The first question was answered by examining verbal utterances across items. Verbatim transcripts from the interviews and the TEMLS forms presented to teachers provided the documents available for analysis. I adopted an analytical approach that allowed patterns of verbalization to emerge from the data through iterative investigation, establishment, and refinement (Coffey & Atkinson, 1996).

The end result of this examination was a descriptive account of patterns of speech. Data analysis began early in the study to allow later data collection efforts to benefit from insights gained in previous interviews and help identify points of stability in the codes assigned to patterns of speech. After each of the first few interviews, I recorded impressions of the main concepts, themes, issues, and questions arising from an interview. My focus at this time was mostly on the think-aloud process, the success with which I was eliciting what I thought would be useful data, and the challenges I faced as a novice with this method as well as a representative

of the university who works on student assessment and testing. I was sensitive to the potential reactivity of the data collection setting. I made self-evaluations related to presenting myself as non-judgmental and neutral, relating to teachers across a wide range of ages, and the extent to which I was facilitating the kind of verbalizations that would eventually allow me to answer my research questions. I also began to note wording issues with the items, misconceptions held by the teachers in relation to educational measurement, the quality of item distractors, and emerging patterns of verbalization (e.g., repetition of item stems). I also recorded these types of impressions after transcribing the first batch of interviews. After exceeding five interviews, my journaling became increasing analytical in orientation; I began analyzing transcripts, and focused mostly on the identification and definition of codes that could be applied to the verbalizations.

Initially, transcripts from the first nine interviews were compiled and reduced to their portions that covered responses to the TEMLS items. The transcript data were re-arranged so that all of the teachers' responses to a single item appeared together (i.e., all of the teachers' responses to item 1, then all of the teachers' responses to item 2, etc.). After an initial reading of the data, memos were recorded, moving from top-to-bottom, characterizing as many sections of verbalization as possible. As commonalities developed around these memos, provisional codes were established. After this establishment, each potentially new instance of a code was compared against previous instances. Codes were added, relabeled, and redefined through this constant act of comparison. After one complete pass through the data, 11 codes were identified. At this time, coded passages were grouped by label and reviewed to see if any new patterns needed to be defined or established patterns re-defined. While consistency was desired within each of the pattern groupings, thematic saturation was of primary concern. Each pattern was judged by its relevance to the research questions.

Following this step of refinement, the codes were examined for thematic commonalities. Existing literature using the think-aloud methodology for the study of reading comprehension among native (Laing & Kamhi, 2002; Pressley & Afflerbach, 1995) and non-native (Anderson, 1991; Gao & Rogers, 2011; Nikolov, 2006) speakers, architecture (Katz, Martinez, Sheehan, & Tatsuoka, 1998), and math (Stylianou, 2002) was consulted for examples of thematic groupings and labels that could be applicable to the present study. Three themes, or *macro-codes*, of response verbalizations were identified—understanding the item, test-management processes, and social and affective strategies. *Understanding the item* involved verbalizations related to identification of the relevant information contained in the item and what is being asked of the respondent. *Test-management processes* captured invocations of various actions to analyze, reason through, solves the items, and justify responses. *Social and affective strategies* embodied indirect statements about the items or respondent. Such verbalizations may be reflective of the social dimension of the think-aloud setting (Nikolov, 2006, p. 28).

The next pass through the data was made with the macro codes in mind. Starting from fresh with data from now 11 interview transcripts, passages of verbalization were coded at the macro level. Again the aim was to capture as much of the verbalization text as possible. Once coding was completed at the macro level, each section of coded text was then broken down to micro-level codes. The initial 11 codes from the first pass were refined to a final set of 12 under the three macro codes. This code structure was confirmed via analysis of the transcript from the 12^{th} and 13^{th} interviews. Analysis suggested no new codes or changes to existing code definitions. One final check was made on all codes, and some re-assignments of text were made to the codes for paraphrasing, re-reading the item, and explaining or reasoning toward an answer. Also the paraphrase code was moved from the *test management processes* macro code to

understanding the item. The final code structure is displayed in Table 4-1. Appendix C displays definitions for each of the macro- and micro-codes.

Understanding the item was comprised of four micro-codes—read item stem, re-read item stem, express confusion/desire more information, and paraphrase. No distinctions were made between items read with or without error (e.g., misreading a word) because such distinctions were irrelevant to the analytical interest in the process undertaken by teachers when engaging the TEMLS knowledge items. Furthermore, detection of errors of omission (i.e., skipping words in the item stem) would have been confounded by the verbalizing skills of the individual teachers. Just because a word was not recited aloud does not mean the teacher did not read the word. The codes for expressing confusion/desiring more information and paraphrasing appear under this macro-code because they involved taking steps to understand what was being asked or stating that full understanding could not be achieved.

Under the *test-management processes* macro-code appeared five micro-codes—read/evaluate response options, strategize/self-direct, explain/reason, rely on prior experience, and apply hypothetical classroom context. Each of these micro-codes captured verbalizations that were directly related to arriving at and feeling confident (to the extent possible) in an answer. For the purposes of deriving codes for mental processes, no distinctions were made between correct and incorrect reasoning. Such distinctions became important in answering the second research question related to support for validation of the TEMLS.

Three micro-codes—self-evaluation/self-encouragement, comment on items, and express uncertainty/admit guessing—comprised the final macro-code, *social and affective strategies*. While several verbalizations within this domain were associated with understanding and solving the items, they were more indirect in nature. That is, they often served as a mental break for the

teacher or were offered as asides. It is unknown to what extent these verbalizations would appear without the social dimension introduced by the presence of the interviewer.

The codes that were derived were intended to be exhaustive in the sense that they captured all utterances that revealed themselves commonly across individuals and were related to the TEMLS items. Not every code was present for each item nor did every teacher verbalize across all codes. Some codes are closely related. One can imagine how a TEMLS respondent could switch between utterances of evaluating response options and explanation/reasoning within a single sentence or two. While each code can be defined in a unique way, not all instances of coded speech were necessarily mutually exclusive. More importantly, as much of the verbalization text related to solving TEMLS items as possible was captured by stable code definitions.

Other sources of data from the interviews support the derived codes. For example, on the test forms, pencil marks reveal cases where teachers evaluated response options. Some teachers used the forms to diagram normal curves and standard deviations to solve a question asking them to judge a student's performance relative to peers when given a mean score and standard deviation. There were also instances of teachers writing out a sequence of numbers to demonstrate a median score. In the retrospective portion of the interview, when asked about the extent to which they drew on past training in assessment when responding to TEMLS items, teachers readily offered personal examples of working with student test data, providing support for the observed pattern of speech related to drawing on prior experience. A similar observation was made in teachers' answers to a question asking the extent to which TEMLS items reflected tasks they carried out with their students' assessment results, thus providing support for the observed code related to applying a hypothetical classroom context

Validation. To answer the second question, the coded interview transcripts were examined alongside information concerning whether the teacher arrived at the correct response or not for each item. The main point of interest in the analysis was to assess the degree to which teachers drew from educational measurement principles to answer the questions (i.e., the extent to which correct answers appeared to reflect understanding of the content being tested). Assessments were made regarding teachers' abilities to understand the questions and correctly interpret what was being asked of them. Verbalizations coded as explaining/reasoning were inspected for their depth and alignment to measurement principles. Differences between the reasoning of teachers providing a correct answer on a given item and those providing an incorrect answer were noted. Also noted were instances in which correct answers were provided even though the teacher admitted guessing or expressed uncertainty about the answer provided or the central topic addressed by the item.

At a basic level, each of the TEMLS knowledge questions was correctly interpreted by every teacher in the study. In no case did a teacher not understand what was being asked or make a fundamental mistake in addressing the question. In general, questions appeared to tap into the intended content knowledge and some degree of understanding of the targeted measurement concept was required to select the correct answer. There were exceptions, however. To correctly answer one question, intended to assess understanding of standard deviations, teachers really only needed to know that the number 700 was greater than 500. A correct response did not require that they understand what a standard deviation is or that a score that is two standard deviations above the mean places the examinee above the 97^{th} percentile. As a result, 12 of 13 teachers answered this question correctly. Revised item distractors could better target understanding of standard deviations and render a more discriminating item.

Another item, asking teachers to interpret a *Z* score of 0, did not capture fundamental understanding of *Z* scores. Based on the responses given in this study, as well as results from pilot administration of the TEMLS (Gotch & French, in press), percentiles are generally well understood by teachers. In this particular question, the correct answer matches a *Z* score of 0 to the 50^{th} percentile. Teachers were able to apply their understanding of percentiles, and work backwards from that understanding to guess what *Z*=0 might mean. Across the 13 teachers, 10 provided a correct answer and every one of them admitted guessing or expressed substantial uncertainty in their solving of the problem. Substantial revision would be required of this item, perhaps moving away from *Z*=0 or tying a *Z* score to a certain scale score, should *Z* scores remain a part of the measurement literacy domain to be addressed by the TEMLS.

A similar observation was made regarding an item that contained the term, "psychometric properties". Only 6 teachers answered this item correctly but among them were 3 who arrived at their answer via guessing, and 9 of the teachers explicitly stated they were not familiar with the meaning of *psychometric*. This item asked teachers what they should do to evaluate the psychometric properties of a test. After expressing unfamiliarity with the terminology of psychometric properties, teachers were able to implicitly translate that term to something akin to "technical quality" in order to proceed with solving the item. What needs to be determined is whether the item is intended to target the action of evaluating a test or the terminology. As the question exists at present, it would probably be best to adopt the former target, and revise the question stem to no longer include the unfamiliar terminology. It might be informative, however, to have an item targeting understanding of the word psychometric, as it is one the measurement community might employ in test manuals and score reports without much consideration for the confusion such a term might cause.

Conclusion

The purpose of this study was to examine the mental processes teachers go through as they respond to knowledge items on the TEMLS instrument, and evaluate from these observations the instrument's ability to support valid inferences of teacher measurement literacy. Codes derived from teacher verbalizations, gathered via think-aloud protocol, show that when responding to TEMLS knowledge items teachers engage in processes that reflect intended interpretations of the items and applications of prior knowledge to solve them. In general, these processes reveal that the TEMLS instrument appropriately captures teachers' understanding of key measurement concepts and their ability to navigate simple hypothetical score interpretation tasks. Exceptions to this general rule do exist, however, and a thorough review of the instrument is justified. Revisions may be major, as highlighted in the previous section, or minor, to improve the alignment between the language used in the items and the language commonly employed by teachers. For example, one TEMLS items presents a student's scores on Reading, Writing, and Math portions of a test, and asks how the teacher would "explain [this student's] profile to her parents". Some teachers remarked that the response options did not reflect how they would communicate to the parents or that the student's "profile" was defined by more than her test scores. The intent of this item is to assess whether or not the teacher can take one student's scores and evaluate them against the proficiency cut scores for the respective subject domains. A revision to improve the item may be as simple as asking the teacher to explain the student's "*score* profile". This particular question was answered correctly by all 13 teachers, so the question stem wording likely would not impact the validity of inferences about a teacher's level of measurement literacy, but revisions such as this one and other minor ones across other items would improve overall performance of the TEMLS, primarily through the reduction of construct-

irrelevant attention and burden of response on the part of the teachers. Note that the need for revisions such as this one and the ones noted above could only be identified via the think-aloud methodology. Thus, these findings lend support to the notion that investigation of response processes is needed more often in test validation work.

Variability was present in the way in which teachers engaged with and solved TEMLS items. Therefore, it was not possible to derive a model of task performance to the extent one is described in relevant literature (Leighton, 2004). While inconsistencies were observed in terms of the flow of responding to items, this study established a suite of codes that can be applied to think-aloud data from the TEMLS knowledge scale. I note that not every code is present among the responses to every item. Items may elicit a lot of verbalization among a few codes, verbalizations across several codes, or few verbalizations at all. The same variance can be observed across participants. Some teachers verbalized a lot within a few codes, others verbalized in more diverse patterns. Future research could employ a larger, purposeful sample to see if there is any relationship between how one solves the problems and one's level of measurement literacy or if the differences observed in the present study are simply idiosyncrasies among how individuals solve multiple-choice items. There is also opportunity to extend the sampling frame beyond the elementary level, as targeted in this study.

Future research should also consider the relationship between interpretation of the TEMLS items and interpretation of standardized test score reports. Several of the general actions that arose in this present analysis could reflect what teachers might go through when receiving score reports and preparing for parent conferences. That is, teachers may read and paraphrase the information presented in order to understand what is being communicated. They may express confusion or desire more information be presented in order to improve clarity. Teachers may rely

on prior experience with score reports and the types of information commonly presented in them to provide explanations of student performance. They may use self-encouragement to work their way through the reports, comment on report content, and express uncertainty about the meaning of various elements of the score reports. Establishing a link between the mental processes undergone in solving TEMLS items and the mental processes engaged in while reviewing score reports would lend further evidence to the ability of the TEMLS to capture meaningful assessments of teacher measurement literacy. Given the success of the methodology to provide a unique type of information in the present study, extension of the think-aloud methodology to score reports to provide evidence of this link is warranted.

Table 4-1

Response process code structure for the TEMLS knowledge scale.

Macro-code	Micro-code	Example(s)
Understanding the item	Read item stem	With a criterion referenced assessment a student's performance is judged in reference to…
	Re-read item stem	With a criterion referenced.
	Express confusion/ desire more information	okay, when comparing standardized test scores, am I comparing test scores from the same test? So, they haven't converted this to a standard score is what I'm wondering. It says raw score it doesn't say standard score.
	Paraphrase	So for both it's a one to nine hundred and 500 is average. Chance in performance, like the kid's tired that day.
Test management processes	Read/evaluate response options	Sam can skip 3rd grade reading lessons. Certainly not that one. No interpretation is possible without further information. I kind of like that one.
	Strategize/self-direct	Let's see if any of these were true. Okay, I need to go back.
	Explain/reason	I'm thinking that the target moved, that's usually what happens in, um, in state testing… …'cause if the other kids scored just as bad or even worse, then it wouldn't fall below the median, 'cause if the kids all scored two points and he scored 33 he's pretty smart.
	Rely on prior experience	When I taught GT, we would give students a placement test to find out kind of where they were at mathematically. I just keep thinking of that MAP test, a Measure of Academic Student Progress, and it does take into account their past performance.
	Apply hypothetical classroom context	He might, he might be a kid I'd test for the gifted program. if this was my student and I go back and the teacher prior said she exceeds standards
Social & affective strategies	Self-evaluation/ self-encouragement	I'm pretty sure I should know this. I mean, I know what a confidence interval is.
	Comment on items	Okay, this is interesting. That's a really good question, really makes you think about it.
	Express uncertainty/ admit guessing	I have no idea I'm just guessing

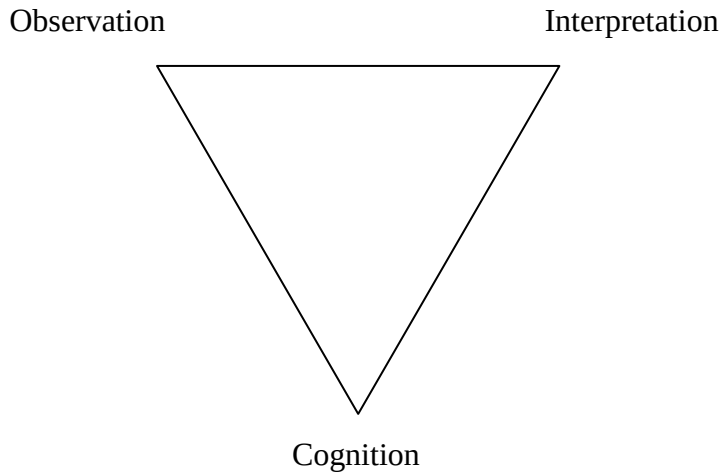

Figure 4-1. The assessment triangle. Adapted from National Research Council, 2001, p. 44.

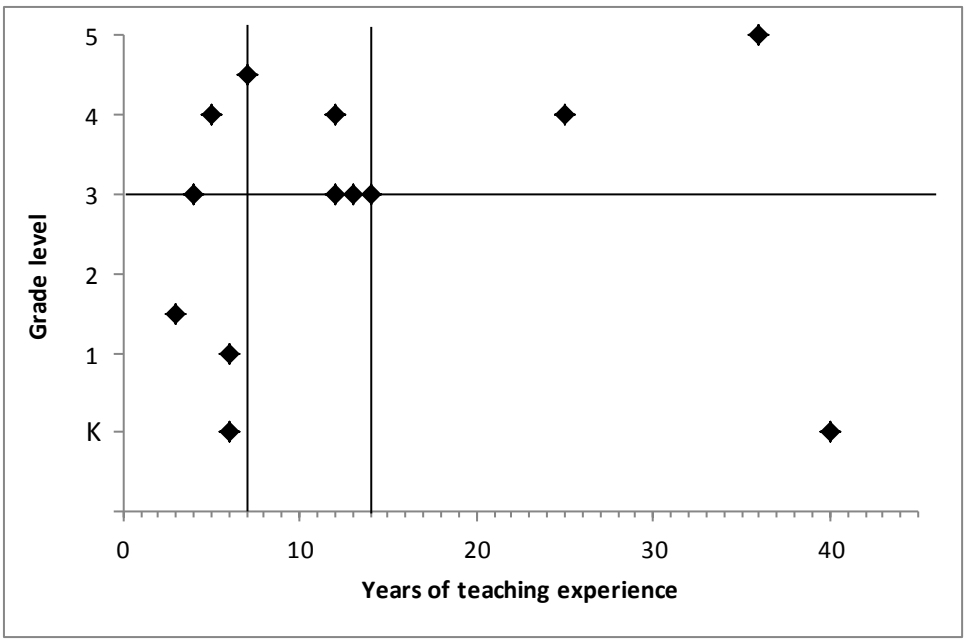

Figure 4-2. Grade level taught at the time of the study by years of teaching experience for teachers participating in the think aloud interviews.

Note: One teacher taught in a 1st/2nd grade combined classroom, and another taught in a 4th/5th combined classroom.

CHAPTER FIVE

CONCLUSION

The overall theme of this dissertation was to investigate teacher competence in working with assessment results through the lens of validity. Validity concerns drawing appropriate inferences from the evidence available. Competence in student assessment therefore perfectly embodies validity with an overarching question of "do teachers draw valid inferences from student performance on assessments?" This dissertation informed this question in two ways. First, a synthesis of the inferences that have been drawn in the literature was attempted. Second, a newly developed instrument aimed at answering the question above was subjected to two disparate forms of evaluation of the evidence the instrument provides.

The need to investigate teacher assessment literacy is ripe. Reflecting trends in society at-large, the U.S. educational system is becoming saturated with data on student achievement and performance. Such data are driving high-stakes decisions, such as student promotion/retention, graduation, and teacher, principal, and school evaluation. To have confidence in these decisions, we need a workforce that is literate in assessment data, where it comes from, what it can tell us and what it cannot. Much of the assessment data in use today, especially those data that are made publicly available, come from standardized tests. Investigations of assessment literacy have traditionally emphasized classroom assessment. While the skills typically associated with this form of assessment, such as aligning in-class assessment to learning objectives and developing reliable and trustworthy grading methods (e.g., via rubrics), with the increasing visibility of standardized test data and its integration in to instruction and evaluation, a particular competence, namely measurement literacy, needs to begin receiving more scholarly attention. Measurement literacy, as defined in this dissertation, concerns the ability to understand and work

with the results of standardized tests. Though one might associate such literacy with only summative forms of assessment, there are potential formative uses of standardized test data available to the teacher or a team of educators. For example, a teacher could use the previous year's standardized test performance of an incoming cohort of students as a piece of information to tailor instructional emphases or strategies for the upcoming year. As another example, in a multi-year program, standardized test results could be used to inform *benchmarks* of progress toward an ultimate goal, thus informing the extent to which the program is playing out according to a part of its design. And perhaps one very valuable skill to have as a measurement literate is to be able to recognize when formative uses of test data are unwarranted (e.g., diagnosis based on unreliable subscales).

Fundamental knowledge of facts does not encompass the totality of literacy. One must consider not only technical skill but also what the individual does with the technical skills at his disposal, how frequently those skills are called upon, and to what ends they are used (Cremin, 1988). Therefore, there is a behavioral component—a kind of internalization of competencies and an orientation to make using assessment results a habit. With a combination of knowledge and self-efficacy content, the Teacher Educational Measurement Scales (TEMLS) were designed to address this component of literacy in measurement. Development of the TEMLS meets the need to have solid measure of the phenomenon of interest before making any defensible claims about teacher competence in measurement tasks and concepts. Data provided by empirical use of the TEMLS may trump easily accessible anecdotal evidence and hopefully drive the science behind training programs and policy efforts.

In the work I presented in this dissertation, I attempted to capture a picture of the work that has been done around teachers' competencies for working with assessments and their results,

and to present evidence suitable for evaluating claims made from a new measure of one form of such competency. The former concerns conclusions the field has drawn about teacher competence in assessment, and was addressed by Chapter 2 of this dissertation. The latter seeks to inform the development and refinement of the TEMLS, as was addressed by Chapters 3 and 4.

The purpose of the study presented in Chapter 2 was to provide a basis for establishing a collective memory (McNeill, 1985) in the area of assessment literacy and to identify gaps and inefficiencies in attention given to the topic. An exhaustive search of readily available literature revealed 65 unique works published in the last two decades, since the concept of assessment literacy was fomented by the *Standards for Teacher Competence in Educational Assessment of Students* and formal definition and promotion. Synthesis of these works showed that we need improvements in measuring assessment literacy and building a cohesive community of scholars to build knowledge on the subject. It raised concerns about what we have learned about teacher competence in assessment, including the measures at our disposal to draw conclusions about such competence. In that sense, the review set the stage for the two other studies presented here. These studies had to do with how well a newly developed instrument of a strand of assessment literacy could serve the field.

The purpose of the study presented in Chapter 3 was to examine the internal structure of a measure of educational measurement self-efficacy, to evaluate the instrument's ability to support inferences regarding this characteristic in teachers and to explore the nature of the measurement self-efficacy construct, one that has not seen extensive attention in the literature. Results were inconsistent for the first aim, thus diminishing our ability to draw conclusions about the latter. A factor model, albeit a complex one, could be fit to TEMLS-SE data with relative success. Some

fit indices, (e.g., CFI, TLI) were high (0.98), but model residuals were large, reflected in the high SRMR value (0.11). The model did not show strict invariance when applied to a second sample.

The purpose of the study presented in Chapter 4 was to examine the mental processes teachers go through as they respond to knowledge items on the TEMLS instrument, and evaluate from these observations the instrument's ability to support valid inferences of teacher measurement literacy. Codes derived from teacher verbalizations, gathered via think aloud protocol, showed that when responding to TEMLS knowledge items teachers engage in processes that reflect what they might go through when receiving score reports and preparing for parent conferences. In general, these processes reveal that the TEMLS instrument appropriately captures teachers' understanding of key measurement concepts and their ability to navigate simple hypothetical score interpretation tasks. Exceptions to this general rule do exist, however, and a thorough review of the instrument is justified.

Validity evidence for the TEMLS

Just as ensuring teachers draw accurate inferences from test results is of high priority, so is making sure that the inferences we draw about teachers' abilities are equally accurate. Therefore the research presented here comprises a validity investigation of the TEMLS content and scores coming from TEMLS responses. Overall, tentative support for the TEMLS was found. We know some TEMLS items are functioning well, while others need revision. The results of the think aloud task point to the need for substantial revisions to items concerning standard deviations, *Z* scores, and evaluating the psychometric properties of a test. The results of the factor analysis suggest revision of an item addressing interpretation of a student's strengths and weaknesses based on mastery level designation and perhaps some reconceptualization of the

family of measurement-related tasks in which teachers might engage. The possibility of unaddressed factors remains.

Claims regarding the general level of a teacher's measurement literacy, or the general literacy level of a population of teachers, can be supported with evidence that most of the TEMLS knowledge items have shown to be correctly interpreted by teachers and accurately reflect understanding of the targeted concepts. Finer distinctions of literacy levels may not be supported, given the item revision needs noted above. Claims may be made of teacher self-efficacy for three families of measurement-related tasks—score/test processing, basic responsibilities, and score transfer. Such claims would be supported by relative good fit of a 3-factor model. Results of the factor analysis, however, would not be able to support use of factor loadings to create factor scores (i.e., via weighted sum), as these loadings were not found to be invariant across different groups of teachers.

Outcomes of the Line of Research

The study of teacher competence in assessment is complex. This is revealed through the sometimes contradictory findings in the literature as well as the validity evaluations presented here. A factor model for measurement self-efficacy involved a complex variable, loading on multiple factors, and several pairs of correlated error terms. This structure could be replicated across random samples of teachers, but the specific values associated with the links between components of this model differed significantly between samples. The literature on the broader notion of assessment literacy displays a lack of cohesion and follow-through. Few have taken it upon themselves to build a research program in the area of study (see Chapter 2). Many findings are contradictory with other studies or represent the lone investigation of a particular research

question. Despite inconsistent results and diminished opportunity for drawing conclusions, we nevertheless have a foundation upon which to build future inquiry.

Once sufficient evidence from a revised version of the TEMLS has been gathered to support central claims about the levels and nature of teachers' measurement literacy, the instrument may be used to guide an extensive line of research. First, efforts will center on establishing baseline levels of measurement literacy within the teaching workforce, including specific topics with high rates of misunderstanding. Associated studies will investigate the relationships between measurement knowledge and self-efficacy with each other as well as with various characteristics of teachers (e.g., years of teaching experience, grade level taught) and communities (e.g., proportions of English-language learners or students meeting accountability benchmarks).

Once baseline levels and associations have been revealed we can begin to draw conclusions about what the field may need to assist teachers in being more successful in interpreting, explaining, and acting upon student score information. I envision a suite of products that can be delivered directly to teachers to help them understand the information that is presented to them. These products may function as a part of a comprehensive professional development effort, providing training to both pre- and in-service teachers. This is also the potential for extension of the work and resources to other groups of individuals, such as students, parents, and school administrators.

An anticipated by-product of this work is a heightened awareness of measurement literacy as a distinct quality. In recognition of this distinctness and the visibility of the concept, we could see an infusion of measurement literacy into the dialogue surrounding assessment systems. As a result, more support could be garnered for professional development resources

targeted toward building measurement capacities. The outputs of the work could also inform various disciplines (e.g., educational psychology) on how to contribute to teacher education. The link between teacher education programs and positive outcomes for either teachers or their students is weak (Patrick, Anderman, Bruening, & Duffin, 2011). The TEMLS would be foundational for any work along these lines to show that these programs build teacher candidates' competencies for assessment, which the National Academy of Education has established as a core competency of prepared teachers (National Academy of Education, 2005).

A better understanding of measurement literacy could also help inform score reporting. A well-articulated model of score reporting takes into account the characteristics of the information recipient (in this case teachers, administrators, parents, and students). At present there are no measurement instruments validated for fulfilling the purpose of providing information about what score report recipients are likely to understand and feel able to act upon.

Empirical work that uses the TEMLS could also contribute to the policy arena. Teacher evaluation is a hot topic in educational policy, particularly in cases where student test scores are being considered. There is doubt, however, about how well student test scores can serve such a purpose, and prominent individuals in the measurement field have advocated for the consideration of teacher assessment literacy as a component in a comprehensive teacher evaluation system (e.g., Brookhart, 2011; Marion, 2011). To achieve status as an effective component of the evaluation process, the field needs to develop nomological networks around assessment literacy that clearly articulate how competencies within this realm impact teacher performance and student learning. It is here that the TEMLS can be utilized to explore the possibilities of such a network.

REFERENCES

Aiken, L. R. (1991). *Psychological testing and assessment* (7th ed.). Boston: Allyn & Bacon.

Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education*, *6*, 241–254. doi:10.1207/s15324818ame0603_4

Alkharusi, H. (2011a). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia - Social and Behavioral Sciences*, *29*, 1614–1624. doi:10.1016/j.sbspro.2011.11.404

Alkharusi, H. (2011b). A logistic regression model predicting assessment literacy among in-service teachers. *Journal of Theory and Practice in Education*, *7*, 280–291.

Alkharusi, H., Kazem, A., & Al-Musawai, A. (2010). Traditional Versus Computer-Mediated Approaches of Teaching Educational Measurement. *Journal of Instructional Psychology*, *37*, 99–111.

Alkharusi, H., Kazem, A. M., & Al-Musawai, A. (2011). Knowledge, skills, and attitudes of preservice and inservice teachers in educational measurement. *Asia-Pacific Journal of Teacher Education*, *39*, 113–123. doi:10.1080/1359866X.2011.560649

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557. doi:10.1037/0021-843X.112.4.545

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

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for Teacher Competence in Educational Assessment of Students* (p. 6).

Anastasi, A. (1993). A century of psychological testing: Origins, problems, and process. *Exploring applied psychology: Origins and critical analyses*, Master lectures in psychology (pp. 9–36). Washington, DC: American Psychological Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*, 460–472. doi:10.1111/j.1540-4781.1991.tb05384.x

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Earlbaum Press.

Arce-Ferrer, A. J., Cab, V. P., & Cisneros-Cohernour, E. J. (2001). Teachers' assessment competencies. Presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Armario, C. (2010, December 7). "Wake-up call": U.S. students trail global leaders. *Associated Press*. Retrieved from http://www.msnbc.msn.com/id/40544897/ns/us_news-life/t/wake-up-call-us-students-trail-global-leaders/

Arter, J. A. (2001). Washington Assessment Professional Development Program evaluation results. Presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Bandalos, D. L. (2004). Can a teacher-led state assessment system work? *Educational Measurement: Issues and Practice, 23*(2), 33–40. doi:10.1111/j.1745-3992.2004.tb00157.x

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Company.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*, 1–26. doi:10.1146/annurev.psych.52.1.1

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents*, Adolescence and Education (pp. 307–337). Greenwich, CT: Information Age Publishing.

Bangert, A. W., & Kelting-Gibson, L. (2006). Teaching principles of assessment literacy through teacher work sample methodology. *Teacher Education and Practice*, *19*, 351–364.

Barr, S. L. (1993). *The state of knowledge of educational assessment in Missouri* (Dissertation). University of Missouri-Columbia, Columbia, MO.

Barry, C. L. (2009, December). Current trends and future directions in educational measurement: Perspective of two presidents. *NCME Newsletter*, *17*(4), 9–11.

Bennett, R. E., & Gitomer, D. H. (2008). *Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support* (No. ETS RM-08-13). Princeton, NJ: Educational Testing Service.

Benson, C. C. (1997). *Assessment and instructional practices of secondary mathematics teachers who participated in project EXTRA: A case study* (Dissertation). University of Missouri-Kansas City, Kansas City, MO.

Betz, J. W. (2009). *Assessment practices in elementary visual art classrooms*. University of Central Florida, Orlando, FL.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. doi:10.1080/0969595980050102

Blume, H. (2011, July 8). Teachers from low-performing schools face stigma on job search. *Los Angeles Times*. Retrieved from latimes.com

Borko, H. (1997). New forms of classroom assessment: Implications for staff development. *Theory into Practice, 36*, 231–238. doi:10.1080/00405849709543773

Borko, H., Davinroy, K. H., Bliem, C. L., & Cumbo, K. B. (2000). Exploring and supporting teacher change: Two third-grade teachers' experiences in a mathematics and literacy staff development project. *The Elementary School Journal, 100*, 273–306.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*, 259–278. doi:10.1016/S0742-051X(96)00024-8

Bosscher, R. J., & Smit, J. H. (1998). Confirmatory factor analysis of the General Self-Efficacy Scale. *Behaviour Research and Therapy, 36*, 339–343. doi:10.1016/S0005-7967(98)00025-4

Braney, B. (2011). *An examination of fourth grade teachers' assessment literacy and its relationship to students' reading achievement* (Dissertation). University of Hartford, Hartford, CT.

Broodhead, M. V. R. (1991). *Training teachers to use the Developmental Assessment Paradigm: A feasibility study* (Dissertation). University of Massachusetts-Amherst, Amherst, MA.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, *10*, 161–180. doi:10.1207/s15324818ame1002_4

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice*, *18*(1), 5–13. doi:10.1111/j.1745-3992.1999.tb00002.x

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5–12. doi:10.1111/j.1745-3992.2003.tb00139.x

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York: Teachers College Press.

Brookhart, S. M. (2011a). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. doi:10.1111/j.1745-3992.2010.00195.x

Brookhart, S. M. (2011b, June). Multi-dimensional set of measures. *NCME Newsletter*, *19*(2), 12.

Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of Project asTTle outputs design* (No. Technical Report 15). University of Auckland, Project asTTle.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Bruce, G. (2004). *The Manitoba Grade Three Assessment: A generative study of teachers' perceptions of the consequences and implications of wide-scale formative assessment on classroom practice* (Master's thesis). University of Manitoba, Winnipeg, MB.

Buck, G., Trauth-Nare, A., & Kaftan, J. (2010). Making Formative Assessment Discernable to Pre-Service Teachers of Science. *Journal of Research in Science Teaching*, *47*, 402–421. doi:10.1002/tea.20344

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Campbell, C., & Evans, J. A. (2000). Investigation of preservice teachers' classroom assessment practices during student teaching. *The Journal of Educational Research*, *93*, 350–355. doi:10.1080/00220670009598729

Center on Education Policy. (2007). State high school exit exams: Working to raise test scores.

Chapman, M. L. (2008). *Assessment literacy and efficacy: Making valid educational decisions* (Dissertation). University of Massachusetts-Amherst, Amherst, MA.

Chappuis, J., Stiggins, R. J., Chappuis, S., & Arter, J. (2012). *Classroom assessment for student learning: Doing it right--using it well* (2nd ed.). Boston: Allyn & Bacon.

Chen, C. C., Greene, P. G., & Crick, A. (1998). Does entrepreneurial self-efficacy distinguish entrepreneurs from managers? *Journal of Business Venturing*, *13*, 295–316. doi:10.1016/S0883-9026(97)00029-3,

Chen, P. (2005). Teacher Candidates' Literacy in Assessment. *Academic Exchange Quarterly*, *62*(5), 62–66.

Chirchir, A. K. (1995). *The relationship between teacher training in measurement and classroom assessment procedures in Kenya's secondary schools* (Master's thesis). University of Ottawa, Ottawa, ON.

Cizek, G. J. (2011, July 25). Cheating on tests and other dumb ideas. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/07/25/37cizek.h30.html

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397–412. doi:10.1177/0013164407310130

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. Thousand Oaks, CA: Sage.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98

Cotton, D., & Gresty, K. (2006). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37*, 45–54. doi:10.1111/j.1467-8535.2005.00521.x

Council of Chief State School Officers, & National Governors Association Center for Best Practices. (2010, June 2). Introduction to the Common Core State Standards. Retrieved from http://www.corestandards.org/assets/ccssi-introduction.pdf

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales: Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42*, 111–131. doi:10.1348/014466503321903544

Cremin, L. A. (1988). *American education: The metropolitan experience 1876-1980*. New York: Harper & Row.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: Sage Publications Inc.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29. doi:10.1037/1082-989X.1.1.16

Daniel, L. G., & King, D. A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research, 91*, 331–344. doi:10.1080/00220679809597563

Daniels, M. H., & Altekruse, M. (1982). The preparation of counselors for assessment. *Measurement and Evaluation in Guidance, 15*, 74–81.

Darker, C. D., & French, D. P. (2009). What sense do people make of a theory of planned behaviour questionnaire?: A think-aloud study. *Journal of Health Psychology, 14*, 861–871. doi:10.1177/1359105309340983

Darling-Hammond, L. (2007). Race, inequality, and educational accountability: The irony of "No Child Left Behind." *Race Ethnicity and Education, 10*, 245–260. doi:10.1080/13613320701503207

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*, 419–438. doi:10.1080/0969594X.2010.516643

Department of Education, Institute of Education Sciences. (2010). Common core of data, 2009-2010. [Data file]. Retrieved from http://nces.ed.gov

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley & Sons.

Edman, E., Gilbreth, S. G., & Wynn, S. (2010). *Implementation of Formative Assessment in the Classroom* (Dissertation). Saint Louis University, Saint Louis, MO.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457. doi:10.1207/S15328007SEM0803_5

Engel, C. (2008). Learning the law. *Journal of Institutional Economics*, *4*, 275–297. doi:10.1017/S1744137408001094

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. doi:10.1037/1082-989X.4.3.272

Fan, Y.-C., Wang, T.-H., & Wang, K.-H. (2011). A web-based model for developing assessment literacy of secondary in-service teachers. *Computers & Education*, *57*, 1727–1740. doi:10.1016/j.compedu.2011.03.006

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York: Psychology Press.

Flemming, M., & Chambers, B. (1984). *Windows on the classroom: A look at teachers' tests*. Captrends. Portland, Oregon: Center for Performance Assessment, Northwest Regional Testing Laboratory.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299. doi:10.1037/1040-3590.7.3.286

Forbes, E. W. (2007). *Improving the knowledge and use of formative assessment: A case study of a model of formative assessment in a K-3 science curriculum* (Dissertation). University of Delaware, Newark, DE.

Freire, P. (1998). Education and conscientizacao (M. B. Ramos, Trans.). In A. M. A. Freire & D. Macedo (Eds.), *The Paulo Freire reader* (pp. 80–110). Continuum. (Reprinted from Education for critical consciousness by P. Freire, 1973, New York: Seabury Press).

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28,* 77–104. doi:10.1177/0265532210364380

Gewertz, C. (2011a, February 23). Common-assessment consortia expand plans. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/02/11/21consortia.h30.html

Gewertz, C. (2011b, August 24). Consortia flesh out visions for common tests. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/08/10/01frameworks.h31.html

Goehring, Jr., H. J. (1973). Course competencies for undergraduate courses in educational tests and measurement. *The Teacher Educator, 9,* 11–20. doi:10.1080/08878737309554546

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7*, 319–355. doi:10.1207/S15328007SEM0703_1

Goodenough, F. L. (1949). *Mental testing: Its history, principles, and applications*. New York: Rinehart & Company.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220. doi:10.1207/s15324818ame1702_3

Goslin, D. A. (1967). *Teachers and testing*. New York: Russell Sage Foundation.

Gotch, C. M., & French, B. F. (in press). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator*.

Gotch, C. M., & French, B. F. (2011). Development and validity evidence for the Teacher Educational Measurement Literacy Scale. Presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Greenhouse, S., & Dillon, S. (2010, March 6). School's shake-up is embraced by the President. *The New York Times*. Retrieved from http://www.nytimes.com/2010/03/07/education/07educ.html

Greenstein, L. (2004). *Finding balance in classroom assessment: High school teachers' knowledge and practice* (Dissertation). Johnson & Wales University, Providence, RI.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research*, *77*, 244–248.

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research*, *79*, 96–100.

Gullickson, A. R. (1986). Teacher Education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*, 347–354. doi:10.1111/j.1745-3984.1986.tb00254.x

Gullickson, A. R., & Hopkins, K. D. (1987). The context of educational measurement instruction for preservice teachers: Professor perspectives. *Educational Measurement: Issues and Practice, 6*(3), 12–16. doi:10.1111/j.1745-3992.1987.tb00501.x

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (No. CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hambrick-Dixon, P. J. (1999). Meeting the challenges to urban school reform: Assessment portfolios for teachers' professional development. Presented at the Annual Meeting of the American Educational Research Association, Montreal, QC, Canada.

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.

Henson, R. K. (2002). From adolescent angst to adulthood: Substantive implications and measurement dilemmas in the development of teacher efficacy research. *Educational Psychologist, 37*, 137–150. doi:10.1207/S15326985EP3703_1

Herold, B. (2011, July 29). Confession of a cheating teacher. *The Philadelphia Public School Notebook*. Retrieved from http://www.edweek.org/ew/articles/2011/07/29/37pnbk_confessions.h30.html

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hickman, J. A. (1991). Does formal training in tests and measurement affect performance on teacher certification exams? *Education, 112*(1), 72.

Hills, J. R. (1977). Coordinators of accountability view teachers' measurement competence. *Florida Journal of Education Research, 19*, 34–44.

Hoover, N. R. (2009). *A descriptive study of teachers' instructional use of student assessment data* (Dissertation). Virginia Commonwealth University, Richmond, VA.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118

Huai, N., Braden, J. P., White, J. L., & Elliott, S. N. (2006). Effect of an internet-based professional development program on teachers' assessment literacy for all students. *Teacher Education and Special Education, 29*, 244–260. doi:10.1177/088840640602900405

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 16–18. doi:10.1111/j.1745-3992.1991.tb00212.x

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Johns, J. L., & VanLeirsburg, P. (1993). The impact of coursework in tests and measurements on assessment literacy. In T. V. Rasinski (Ed.), *Inquiries in literacy learning and instruction: the fifteenth yearbook of the College Reading Association*. Ann Arbor, MI: College Reading Association.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL*. Chicago: Scientific Software International, Inc.

Kane, M. T. (2006). Validation. *Educational measurement*, American Council on Education / Praeger Series on Higher Education (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–773. doi:10.1126/science.1199327

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics*, *24*, 254–278. doi:10.3102/10769986023003254

Kershaw, I. (1993). *Ohio vocational education teachers' perceived use of student assessment information in educational decision-making* (Dissertation). The Ohio State University, Columbus, OH.

King, J. D. (2010). *Criterion-referenced assessment literacy of educators* (Dissertation). University of Southern Mississippi, Hattiesburg, MS.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, *22,* 255–276. doi:10.1080/10476210.2011.593164

Laing, S. P., & Kamhi, A. G. (2002). The use of think-aloud protocols to compare inferencing abilities in average and below-average readers. *Journal of Learing Disabilities*, *35*, 437–448. doi:10.1177/00222194020350050401

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15. doi:10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom-versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice, 17*, 7–21. doi:10.1080/09695940903565362

Lingard, B., Mills, M., & Hayes, D. (2006). Enabling and aligning assessment for learning: Some research and policy lessons from Queensland. *International Studies in Sociology of Education, 16*, 83–103. doi:10.1080/09620210600849778

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson.

Lomax, R. G. (1996). On becoming assessment literate: an initial look at preservice teachers' beliefs and practices. *The Teacher Educator, 31*, 292–303. doi:10.1080/08878739609555122

Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. *Developing early literacy: Report of the National Early Literacy Panel* (pp. 55–106). Washington, DC: National Institute for Literacy. Retrieved from http://lincs.ed.gov/publications/pdf/NELPReport09.pdf

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*, 73–79. doi:10.1027/1614-2241.4.2.73

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26–32. doi:10.1111/j.1745-3992.2004.tb00156.x

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. doi:10.1037/1082-989X.4.1.84

MacInnis, L., & Lambert, L. (2011, September 24). Obama links education reform to economic recovery. *Reuters*. Retrieved from http://www.reuters.com/article/2011/09/24/usa-education-obama-idUSS1E78M22620110924

Maclellan, E. (2004). Initial knowledge states about assessment: novice teachers' conceptualisations. *Teaching and Teacher Education*, *20*, 523–535. doi:10.1016/j.tate.2004.04.008

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29–46. doi:10.1080/01619568809538611

Marion, S. F. (2011, June). The challenges of incorporating student performance results from "non-tested" grades and subjects into educator effectiveness determinations. *NCME Newsletter*, *19*(2), 15–16.

Mayo, S. T. (1967). *Pre-service preparation of teachers in educational measurement* (Final Report. No. (ERIC Document Reproduction Service No. ED 021 784)). Chicago: Loyola University.

Mayo, S. T. (1970). *Trends in the teaching of the first course in measurement* (National Council on Measurement in Education Symposium Paper No. (ERIC Document Reproduction Service No. ED 047 007)). Chicago: Loyola University.

Mazor, K. M., Canavan, C., Farrell, M., Margolis, M. J., & Clauser, B. E. (2008). Collecting validity evidence for an assessment of professionalism: Findings from think-aloud interviews. *Academic Medicine*, *83*(10), S9–S12. doi:10.1097/ACM.0b013e318183e329

Mazzie, D. D. (2008). *The effects of professional development related to classroom assessment on student achievement in science* (Dissertation). University of South Carolina, Columbia, SC.

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality, 60,* 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x

McMillan, J. H. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction* (4th ed.). Boston: Pearson.

McMorris, R. F., & Boothroyd, R. A. (1993). Tests that teachers build: An analysis of classroom tests in science and mathematics. *Applied Measurement in Education, 6,* 321–342. doi:10.1207/s15324818ame0604_5

McNeill, W. H. (1985). Why study history. American Historical Association. Retrieved from http://www.historians.org/pubs/archives/whmcneillwhystudyhistory.htm

Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment. *Mid-Western Educational Researcher, 13*(4), 29–35.

Mertler, C. A. (2003). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? Presented at the Annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33,* 49–64.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12,* 101–113. doi:10.1177/1365480209105575

Mertler, C. A., & Campbell, C. (2005). Measuring teachers' knowledge & application of classroom assessment concepts: Development of the Assessment Literacy Inventory. Presented at the Annual meeting of the American Educational Research Association, Montreal, QC, Canada.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis* (2nd ed.). Thousand Oaks, CA: Sage.

National Academy of Education. (2005). *A good teacher in every classroom: Preparing the highly qualified teachers our children deserve*. San Francisco, CA: Wiley.

National Council on Education and the Disciplines. (2001). The case for quantitative literacy. In L. A. Steen (Ed.), *Mathematics and democracy: The case for quantitative literacy* (pp. 1–22). Princeton, NJ: Author.

National Research Council. (2001). *Knowing what students know: The Science and design of educational assessment*. Washington, DC: National Academy Press.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organization Research Methods, 6*, 328–362. doi:10.1177/1094428103254673

Ng, T. W. H., Sorensen, K. L., & Eby, L. T. (2006). Locus of control at work: A meta-analysis. *Journal of Organzational Behavior*, *27*, 1057–1087. doi:10.1002/job.416

Nikolov, M. (2006). Test-taking strategies of 12- and 13-year-old Hungarian learners of EFL: Why whales have migraines. *Language Learning, 56*, 1–51. doi:10.1111/j.0023-8333.2006.00341.x

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84,* 231–259. doi:10.1037/0033-295X.84.3.231

Noll, V. H. (1955). Requirements in educational measurement for prospective teachers. *School and Society, 82,* 88–90.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

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. doi:10.1111/j.1745-3992.1991.tb00173.x

O'Sullivan, R. G., & Johnson, R. L. (1993). Using Performance Assessments To Measure Teachers' Competence in Classroom Assessment. Presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Ogan-Bekiroglu, F. (2009). Assessing assessment: Examination of pre-service physics teachers' attitudes towards assessment and factors affecting their attitudes. *International Journal of Science Education, 31,* 1–39. doi:10.1080/09500690701630448

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 151,* 53–79. doi:10.1016/S0377-2217(02)00578-7

Otterman, S. (2011, July 4). Union shifts position on teacher evaluations. *The New York Times*. Retrieved from http://www.nytimes.com/

Page, K., & Uncles, M. (2004). Consumer knowledge of the World Wide Web: Conceptualization and measurement. *Psychology and Marketing, 21,* 573–591. doi:10.1002/mar.20023

Pajares, F., Hartley, J., & Valiante, G. (2001). Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development, 33,* 214–221.

Patrick, H., Anderman, L. H., Bruening, P. S., & Duffin, L. C. (2011). The role of educational psychology in teacher education: Three challenges for educational psychologists. *Educational Psychologist, 46,* 71–83. doi:10.1080/00461520.2011.538648

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher, 39,* 110–119. doi:10.3102/0013189X10363007

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief*. The Aspen Institute.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Upper Saddle River, NJ: Merrill Prentice Hall.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10–12. doi:10.1111/j.1745-3992.1993.tb00548.x

Popham, W. J. (2003). Seeking redemption for our psychometric sin. *Educational Measurement: Issues and Practice, 22*(1), 45–48. doi:10.1111/j.1745-3992.2003.tb00117.x

Popham, W. J. (2005). *Classroom assessment: What teachers need to know* (4th ed.). Boston: Allyn & Bacon.

Popham, W. J. (2006). Needed: A dose of assessment literacy. *Educational Leadership*, *63*(6), 84–85.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*, 4–11. doi:10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*, 265–273. doi:10.1080/08878730.2011.605048

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322–338. doi:10.1037/a0014996

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Quilter, S. M., & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, *36*, 115–131. doi:10.1080/08878730009555257

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, *17*, 1–24. doi:10.1207/s15324818ame1701_1

Roduta Roberts, M., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*(3), 25–38. doi:10.1111/j.1745-3992.2010.00181.x

Roeder, H. H. (1972). Are today's teachers prepared to use tests? *Peabody Journal of Education*, *49*, 239–240. doi:10.1080/01619567209537858

Roeder, H. H. (1973). Education curricula: Your final grade is F. *Journal of Educational Measurement*, *10*, 141–143. doi:10.1111/j.1745-3984.1973.tb00791.x

Roediger, III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Ross, J. A., Cousins, J. B., & Gadalla, T. (1996). Within-teacher predictors of teacher efficacy. *Teaching and Teacher Education, 12,* 385–400. doi:10.1016/0742-051X(95)00046-M

Rotter, J. B. (1966). Generalizes expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80,* 1–28. doi:10.1037/h0092976

Rudner, L. M. (Ed.). (1980). *Testing in our schools*. Washington, DC: National Institute of Education.

SAS Institute, Inc. (2008). *SAS*. Cary, NC: SAS Institute, Inc.

Sato, M., Chung, R. R., & Darling-Hammond, L. (2008). Improving Teachers' Assessment Practices Through Professional Development: The Case of National Board Certification. *American Educational Research Journal, 45,* 669–700. doi:10.3102/0002831208316955

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514. doi:10.1080/09500690701630448

Sawchuk, S. (2011, July 15). D.C. evaluations target hundreds for firing or bonuses. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/07/15/37dismiss.h30.html?

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. doi:10.1037/1082-989X.7.2.147

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3–6. doi:10.1111/j.1745-3992.1991.tb00170.x

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice, 32*, 118–126. doi:10.1080/00405849309543585

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57–63. doi:10.1177/002248718703800312

Schafer, W. D., & Mufson, D. (1993, April). *Assessment literacy for school counselors*. Presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Schmitt, V. L. (2007). *The quality of teacher-developed rubrics for assessing student performance in the classroom* (Dissertation). University of Kansas, Lawrence, KS.

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). New York: Routledge.

Scott, S., Webber, C. F., Aitken, N., & Lupart, J. (2011). Developing teachers' knowledge, beliefs, and expertise: Findings from the Alberta Student Assessment Study. *The Educational Forum, 75*(2), 96–113. doi:10.1080/00131725.2011.552594

Scribner-Maclean, M. (1999). *Exploration of the assessment practices of elementary teachers using science kits* (Dissertation). University of Massachusetts-Lowell, Lowell, MA.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Boston: Houghton Mifflin.

Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention* (4th ed.). New York: Guilford.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22,* 371–391. doi:10.1007/s10972-011-9231-6

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing.

Siwatu, K. O. (2007). Preservice teachers' culturally responsive teaching self-efficacy and outcome expectancy beliefs. *Teaching and Teacher Education, 23,* 1086–1101. doi:10.1016/j.tate.2006.07.011

State of Washington Superintendent of Public Instruction. (2008a, September 11). Personnel by major position and gender for school year 2007-2008. Retrieved from http://www.k12.wa.us/DataAdmin/pubdocs/personnel/StaffGenderREPORT07-08.pdf

State of Washington Superintendent of Public Instruction. (2008b, September 12). Personnel by major position and ethnic for school year 2007-2008. Retrieved from http://www.k12.wa.us/DataAdmin/pubdocs/personnel/StaffEthnicREPORT07-08%20.pdf

State of Washington Superintendent of Public Instruction. (2011). Demographic Information by School. Retrieved from http://reportcard.ospi.k12.wa.us/DataDownload.aspx?schoolId=1&OrgTypeId=1&reportLevel=State&orgLinkId=

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72,* 534–539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77,* 238–245.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, *18*(1), 23–27. doi:10.1111/j.1745-3992.1999.tb00004.x

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

Stiggins, R. J. (2007). Conquering the formative assessment frontier. *Formative classroom assessment: Theory into practice* (pp. 8–28). New York: Teachers College Press.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22*, 271–286. doi:10.1111/j.1745-3984.1985.tb01064.x

Stiggins, R. J., & Chappuis, J. (2006). What a difference a word makes: Assessment for learning rathern than assessment of learning helps students succeed. *Journal of Staff Development*, *27*(1), 10–14.

Stiggins, R. J., & Chappuis, J. (2008). Enhancing student learning. *District Administration*, *44*(1), 42–44.

Stiggins, R. J., & Conklin, N. F. (1988). *Teacher training in assessment*. Portland, OR: Northwest Regional Educational Laboratory.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice*, *5*(2), 5–17. doi:10.1111/j.1745-3992.1986.tb00473.x

Stylianou, D. A. (2002). On the interaction of visualization and analysis: The negotiation of a visual representation in expert problem solving. *Journal of Mathematical Behavior*, *21*, 303–317. doi:10.1016/S0732-3123(02)00131-1

Taylor, C. S., & Nolen, S. B. (1995). A question of validity: A model for teacher training in assessment. Presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The contemporary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *92,* 413–425. doi:I0.1037//0022-0663.92.3.413

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17,* 783–805. doi:10.1016/S0742-051X(01)00036-1

U.S. Department of Education. (2010, March). A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act. Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

United States Congress. (2002). No Child Left Behind Act of 2001. Public Law 107-110. 107th Cong., pp. 1425-2095.

Vanden Berk, E. J. (2005). *Improving the evaluation of students through teacher training: An investigation of the utility of the Student Evaluation Standards* (Dissertation). University of Iowa, Iowa City, IA.

VanLeirsburg, P., & Johns, J. L. (1991). *Assessment literacy: Perceptions of preservice and inservice teachers regarding ethical considerations of standardized testing procedures* (No. Literacy Research Report No. 12). DeKalb, IL: Curriculum and Instruction Reading Clinic.

Vaznis, J. (2011, June 29). Student scores to be key factor in teacher evaluations. *The Boston Globe*. Retrieved from http://boston.com

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education*, *30*, 749–770.

Volante, L., & Melahn, C. (2005). Promoting assessment literacy in teachers: Lessons from the Hawaii School Assessment Liaison Program. *Pacific Educational Research Journal*, *13*, 19–34.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, *36*(4), 301–335.

Wang, T.-H., Wang, K.-H., & Huang, S.-C. (2008). Designing a Web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education*, *51*, 448–462. doi:10.1016/j.compedu.2007.06.010

Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy-value theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 55–75). New York: Routledge.

Williams, L., & Rink, J. (2003). Teacher competency using observational scoring rubrics. *Journal of Teaching in Physical Education*, *22*, 552–572.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*, 37–42. doi:10.1177/002248719104200106

Woolfolk, A. E. (1995). *Educational psychology* (6th ed.). Boston: Allyn & Bacon.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Mahwah, NJ: Lawrence Erlbaum.

Zapata-Rivera, D., Underwood, J. S., & Bauer, M. (2005). Advanced reporting systems in assessment environments. *Learner modelling for reflection, to support learner control, metacognition and improved communication* (pp. 23–31). Presented at the 12th International Conference on Artificial Intelligence in Education, Amsterdam, the Netherlands.

Zhang, Z. (1996). Teacher assessment competency: A Rasch model analysis. Presented at the Annual Meeting of the American Educational Research Association, New York.

Zhang, Z., & Burry-Stock, J. (1995). A multivariate analysis of teachers' perceived assessment competency as a function of measurement training and years of teaching. Presented at the Annual Meeting of the Mid-South Educational Research Association, Biloxi, MS.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*, 323–342. doi:10.1207/S15324818AME1604_4

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age Publishing.

Zwick, R., Sklar, J. C., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice, 27*(2), 14–27. doi:10.1111/j.1745-3992.2008.00119.x

APPENDIX A

PUBLISHED WORKS INCLUDED IN THE SYSTEMATIC REVIEW OF LITERATURE

Study	Author affiliations	Format	Measures	Sample (teachers only)
Alkharusi, 2011a	Sultan Qaboos University (Oman)	Conference Paper	knowledge test (TALQ), translated to Arabic	pre-service teachers
Alkharusi, 2011b	Sultan Qaboos University (Oman)	Article	knowledge test (TALQ), translated to Arabic	in-service middle school teachers
Alkharusi et al., 2010	Sultan Qaboos University (Oman)	Article	knowledge test (TALQ), translated to Arabic	pre-service teachers
Alkharusi et al., 2011	Sultan Qaboos University (Oman)	Article	knowledge test (TALQ), translated to Arabic	pre-service and in-service teachers
Arce-Ferrer et al., 2001	Universidad Autonoma de Yucatan; University of Illinois Urbana-Champaign	Conference Paper	self-report questionnaire, guided by Standards (Spanish)	pre-service and in-service teachers
Arter, 2001	Assessment Training Institute	Conference Paper	teacher work samples	in-service teachers
Bandalos, 2004	University of Georgia	Article	survey and focus groups; self-report on learning gains	in-service teachers
Bangert & Kelting-Gibson, 2006	Montana State University	Article	teacher work samples	pre-service teachers
Barr, 1993	University of Missouri-Columbia	Dissertation	instrument based on Standards	in-service teachers
Benson, 1997	University of Missouri-Kansas City	Dissertation	knowledge test (TALQ); portfolios; final "exam" from training	in-service secondary math teachers
Borko, 1997	University of Colorado-Boulder	Article	teacher assessment ideas and practices; interviews	in-service elementary teachers

Borko et al., 1997	University of Colorado-Boulder	Article	workshop artifacts; interviews	in-service elementary teachers
Borko et al., 2000	University of Colorado-Boulder	Article	workshop artifacts; interviews; observations of teacher practice	2 veteran in-service elementary teachers
Braney, 2011	University of Hartford	Dissertation	knowledge test (API)	in-service elementary teachers
Broodhead, 1991	University of Massachusetts-Amherst	Dissertation	knowledge test; questionnaire	in-service teachers
Bruce, 2004	University of Manitoba (Canada)	Dissertation	interviews	in-service elementary teachers
Buck, et al., 2010	Indiana University; University of Nebraska-Lincoln	Article	pre/post questionnaires; course artifacts; focus groups	pre-service teachers
Campbell & Evans, 2000	Northern Illinois University; Peoria (IL) Public Schools	Article	lesson plans evaluated against a set of criteria	pre-service teachers
Chapman, 2008	University of Massachusetts-Amherst	Dissertation	questionnaire	in-service secondary teachers
Chen, 2005	Hunter College	Article	knowledge test (TALQ)	pre-service teachers
Chirchir, 1995	University of Ottawa (Canada)	Dissertation	questionnaire; interviews	in-service secondary teachers
Daniel & King, 1998	Southern Mississippi; Learning Solutions	Article	T/F test	in-service teachers
DeLuca & Klinger, 2010	Univerity of South Florida; Queen's University (Canada)	Article	questionnaire	pre-service teachers
Edman et al., 2010	St. Louis University	Dissertation	questionnaire	in-service teachers
Fan et al., 2011	National Changhua University of Education; National	Article	AKT	in-service secondary teachers

	Hsinchu University of Education (Taiwan)			
Forbes, 2007	University of Delaware	Dissertation	survey; direct observations; interviews	in-service elementary teachers
Greenstein, 2004	Johnson & Wales University	Dissertation	survey; interviews	in-service secondary teachers
Hambrick-Dixon, 1999	Hunter College	Conference Paper	Evaluation and Assessment of Children Self-Assessment Questionnaire	pre-service and in-service teachers
Hickman, 1991	University of Texas-Autsin	Article	ExCET certification exam	pre-service teachers
Hoover, 2009	Virginia Commonwealth University	Dissertation	survey	in-service elementary and secondary teachers
Huai et al., 2006	University of Wisconsin-Eau Claire	Article	T/F, short-answer test	in-service teachers, support staff, and administrators
Johns & VanLeirsburg, 1993	Northern Illinois University; Elgin Public Schools (IL)	Book Chapter	survey based on Standards	in-service and pre-service teachers
Kershaw, 1993	Ohio State University	Dissertation	self-report questionnaire, guided by Standards	in-service vocational ed secondary teachers
King, 2010	University of Southern Mississippi	Dissertation	test	in-service teachers
Koh, 2011	Nanyang Technical University (Singapore)	Article	teacher assessment examples evaluated against a set of criteria	in-service elementary teachers

Lingard et al., 2006	University of Edinburgh (UK), University of Queensland (Australia), Griffith University (Australia)	Article	observations of teacher practice	in-service elementary and secondary teachers
Lomax, 1996	Northern Illinois University	Article	student journals; teacher example assessments; interviews	pre-service elementary teachers (with some moving to in-service through the course of the study)
Lukin et al., 2004	Lincoln Public Schools; University of Georgia; Edgewood College	Article	self-report questionnaires	in-service teachers (preliminary investigation with pre-service teachers)
Maclellan, 2004	University of Strathclyde (UK)	Article	teacher-produced written text about assessment	teachers just at the point of licensure application
Mazzie, 2008	University of South Carolina	Dissertation	knowledge test	in-service teachers
McMorris & Boothroyd, 1993	SUNY-Albany; NY State Office of Mental Health	Article	knowledge tests; samples of teacher-made tests	in-service middle school teachers
Mertler, 2000	Bowling Green State University	Article	survey	in-service elementary and secondary teachers
Mertler, 2003	Bowling Green State University	Conference Paper	knowledge test (CALI)	pre-service and in-service teachers
Mertler, 2004	Bowling Green State University (assumed)	Article	knowledge test (CALI)	pre-service and in-service secondary teachers
Mertler, 2009	University of West Georgia	Article	knowledge test (CALI)	in-service elementary teachers

Mertler & Campbell, 2005	Bowling Green State University; Northern Illinois University	Conference Paper	knowledge test (CALI)	pre-service teachers
O'Sullivan & Johnson, 1993	University of North Carolina-Greensboro	Conference Paper	knowledge test (TALQ)	teachers in a graduate measurement course
Plake et al., 1993	University of Nebraska-Lincoln; University of South Dakota	Article	knowledge test (TALQ)	in-service teachers
Quilter & Gallini, 2000	University of South Carolina	Article	questionnaire	in-service elementary and secondary teachers
Sato et al., 2008	Unniversity of Minnesota-TC; Stanford University	Article	teacher assessment plans and samples; interviews; surveys	inservice middle and secondary teachers
Schmitt, 2007	University of Kansas	Dissertation	teacher assessment samples	in-service teachers
Scott et al., 2011	University of Calgary; Thompson Rivers University; University of Lethbridge; University of Alberta (Canada)	Article	questionnaire; interviews	in-service teachers
Scribner-Maclean, 1999	University of Massachusetts-Lowell	Dissertation	knowledge test (TALQ); observations; interviews	in-service elementary teachers
Siegel & Wissehr, 2011	University of Missouri-Columbia; University of Arkansas	Article	teaching philosophy essay; reflective journal; teacher work sample	pre-service secondary teachers
Taylor & Nolen, 1995	University of Washington	Conference Paper	questionnaire	pre-service elementary and secondary teachers
Vanden Berk, 2005	University of Iowa	Dissertation	questionnaire	in-service secondary teachers

VanLeirsburg & Johns, 1991	Elgin Public Schools; Northern Illinois University	Report	questionnaire developed from guidelines of test experts (pre-Standards)	pre-service and in-service teachers
Volante & Fazio, 2007	Brock University (Canada)	Article	questionnaire	pre-service teachers
Volante & Melahn, 2005	Brock University (Canada)	Article	questionnaire	in-service teachers
Wang et al., 2008	National Changhua University of Education; National Hsinchu University of Education (Taiwan)	Article	knowledge tests	pre-service teachers
Williams & Rink, 2003	University of South Carolina	Article	teacher assessment examples evaluated against a set of criteria	in-service teachers
Zhang, 1996	Virginia Military Institute	Conference Paper	knowledge test (API)	in-service teachers
Zhang & Burry-Stock, 1995	University of Alabama	Conference Paper	knowledge test (API)	in-service teachers
Zhang & Burry-Stock, 2003	Fairfax County Public Schools; University of Alabama	Article	knowledge test (API)	in-service teachers
Zwick et al., 2008	University of California-Santa Barbara; Cal Poly-San Luis Obispo	Article	knowledge test	pre-service and in-service teachers

APPENDIX B

THINK-ALOUD INTERVIEW PROTOCOL

Introduction

What we would like you to do today is to complete a part of an assessment, and just talk through what you are thinking as you respond. In other words, we would like you to think aloud. The study will work best if you say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. I will remind you to keep talking if you are silent for any long period of time. If you feel nervous about this task, know that it is the instrument that is under evaluation. We are not evaluating your knowledge, attitudes, or capabilities as a teacher.

First I will have you complete a few practice questions. We will do them one-at-a-time, and I'll provide you feedback after each one. If you have any questions about the process after completing these items, you may ask them at that time.

Practice item #1

Circle 2/3 of the 12 ◊s.

◊ ◊ ◊ ◊ ◊ ◊

◊ ◊ ◊ ◊ ◊ ◊

Practice item #2

How well do you believe you can take a picture and email it to a friend using a smartphone (iPhone, Blackberry, Android)? (Please circle your answer.)

1	2	3	4	5	6	7
Not at all well			Moderately well			Very well

Practice item #3

The main purpose of research is to

a) be a phenomenon
b) cause a phenomenon
c) investigate what caused a phenomenon
d) to prevent a phenomenon

Do you have any questions about what you are being asked to do today?

Let's begin. Please complete the questions in this packet, and say your thoughts out loud as you respond to each one.
Present TEMLS items.

Follow-up questions

1) When answering the questions, how much did you use any recent training you may have had in statistics, large-scale assessment, or educational measurement?
2) How reflective were the questions of the kinds of tasks you carry out related to your students' test results?
3) Thinking about standardized test scores, what is the one concept that is hardest to understand or explain?
4) What would be one suggestion you have for helping teachers with the use of score reports?
5) Do you have anything else you'd like to add about anything I've asked you?
6) Was there anything I should have asked you about that I didn't?

APPENDIX C

CODE DEFINITIONS FROM THINK-ALOUD ANALYSIS

Understanding the item

Verbalizations related to identification of the relevant information contained in the item and what is being asked of the respondent

Read item stem

Recitation of the item stem in part or entirely; item does not have to be read verbatim or with 100% accuracy

Re-read item stem

Subsequent recitation of the entire item stem or any portion of it; may include isolation of a key word or phrase

Express confusion/desire more information

Statement of a lack of understanding of item stem content or that more information is needed to feel confident answering the question

Paraphrase

Restatement of the item stem in an effort to understand what is being asked

Test-management processes

Invocations of various actions to analyze, reason through, solves the items, and justify responses

Read/evaluate response options

Recitation of a response option; may be accompanied by a basic evaluation of whether the response is a viable answer choice

Strategize/self-direct

Narration directed at steps needed to solve the problem

Explain/reason

Verbalization expositing measurement principles, providing more rationale for why a response option is or is not a viable answer choice, or justifying an answer choice after the selection has been made

Rely on prior experience

Invocation of concrete experience with previous experience with receiving or using standardized test results or verbalization related to learning the measurement principle in a higher education setting

Apply hypothetical classroom context

Invocation of a scenario where the item is considered in the context of the teacher's professional duties

Social and affective strategies

Indirect statements about the items or respondent. Such verbalizations may be reflective of the social dimension of the think-aloud setting

Self-evaluation/ self-encouragement

Statement related to the ability to solve an item

Comment on items

Statement about an item's content, difficulty, or any other characteristics that are not related to trying to understand the item

Express uncertainty/ admit guessing

Disclosure regarding the unfamiliarity of a word or phrase or admission that an answer choice is being made without a clear rationale

APPENDIX D

HUMAN SUBJECTS FORMS

MEMORANDUM
TO: Brian French and Chad Gotch,
FROM: Patrick Conner, Office of Research Assurances (3005)
DATE: 4/28/2010
SUBJECT: Certification of Exemption, IRB Number 11361

Based on the Exemption Determination Application submitted for the study titled "Educational Measurement Self-Efficacy, Knowledge, and Attitudes: What Teachers Know and What they Want to Know," and assigned IRB # 11361, the WSU Office of Research Assurances has determined that the study satisfies the criteria for Exempt Research at 45 CFR 46.101(b)(2).

This study may be conducted according to the protocol described in the Application without further review by the IRB.

It is important to note that certification of exemption is NOT approval by the IRB. You may not include the statement that the WSU IRB has reviewed and approved the study for human subject participation. Remove all statements of IRB Approval and IRB contact information from study materials that will be disseminated to participants.

This certification is valid only for the study protocol as it was submitted to the ORA. Studies certified as Exempt are not subject to continuing review (this Certification does not expire). If any changes are made to the study protocol, you must submit the changes to the ORA for determination that the study remains Exempt before implementing the changes (The Request for Amendment form is available online at http://www.irb.wsu.edu/documents/forms/rtf/Amendment_Request.rtf).

Exempt certification does NOT relieve the investigator from the responsibility of providing continuing attention to protection of human subjects participating in the study and adherence to ethical standards for research involving human participants.

In accordance with WSU Business Policies and Procedures Manual (BPPM), this Certification of Exemption, a copy of the Exemption Determination Application identified by this certification and all materials related to data collection, analysis or reporting must be retained by the Principal Investigator for THREE (3) years following completion of the project (BPPM 90.01).

Washington State University is covered under Human Subjects Assurance Number FWA00002946 which is on file with the Office for Human Research Protections (OHRP).

Review Type: New
Review Category: Exempt
Date Received: 4/7/2010
Exemption Category: 45 CFR 46.101 (b)(2)
OGRD No.: N/A
Funding Agency: N/A

MEMORANDUM
TO: Brian French and Chad Gotch,
FROM: Patrick Conner, Office of Research Assurances (3005)
DATE: 9/2/2011
SUBJECT: Certification of Exemption, IRB Number 12113

Based on the Exemption Determination Application submitted for the study titled "Response Process Validation Evidence of the Teacher Educational Measurement Literacy Scales," and assigned IRB # 12113, the WSU Office of Research Assurances has determined that the study satisfies the criteria for Exempt Research at 45 CFR 46.101(b)(2).

This study may be conducted according to the protocol described in the Application without further review by the IRB.

It is important to note that certification of exemption is NOT approval by the IRB. You may not include the statement that the WSU IRB has reviewed and approved the study for human subject participation. Remove all statements of IRB Approval and IRB contact information from study materials that will be disseminated to participants.

This certification is valid only for the study protocol as it was submitted to the ORA. Studies certified as Exempt are not subject to continuing review (this Certification does not expire). If any changes are made to the study protocol, you must submit the changes to the ORA for determination that the study remains Exempt before implementing the changes (The Request for Amendment form is available online at http://www.irb.wsu.edu/documents/forms/rtf/Amendment_Request.rtf).

Exempt certification does NOT relieve the investigator from the responsibility of providing continuing attention to protection of human subjects participating in the study and adherence to ethical standards for research involving human participants.

In accordance with WSU Business Policies and Procedures Manual (BPPM), this Certification of Exemption, a copy of the Exemption Determination Application identified by this certification and all materials related to data collection, analysis or reporting must be retained by the Principal Investigator for THREE (3) years following completion of the project (BPPM 90.01).

Washington State University is covered under Human Subjects Assurance Number FWA00002946 which is on file with the Office for Human Research Protections (OHRP).

Review Type: New
Review Category: Exempt
Date Received: 8/19/2011
Exemption Category: 45 CFR 46.101 (b)(2)
OGRD No.: N/A
Funding Agency: N/A

APPENDIX E

CURRICULUM VITA

CHAD M. GOTCH

Dept. of Educational Leadership & Counseling Psychology
Cleveland Hall
Washington State University
Pullman, WA 99164-2136

Phone: (509) 335-8394
Fax: (509) 335-6961
cgotch@wsu.edu

EDUCATION

Ph.D. Education, Anticipated 2012
Washington State University, Pullman, WA
Area of specialization: Educational psychology (measurement and research methods)
Dissertation: *Development of and validation evidence for a teacher educational measurement literacy instrument*

M.S. Resource Recreation & Tourism (Conservation Social Sciences), 2002
University of Idaho, Moscow, ID
Area of specialization: Environmental education

B.S. Natural Resource Recreation, 2001
Virginia Polytechnic Institute and State University, Blacksburg, VA
Minor: Forestry

PROFESSIONAL EXPERIENCE

Graduate Research Assistant, 2008 – present
Dept. of Educational Leadership & Counseling Psychology, Washington State University, Pullman, WA

I have worked on the re-standardization process of two national basic skills tests, including application of classical and modern test theory to tasks of scaling, norming, and validation; development of scoring sheets and score tables; and assistance with writing the technical manuals. For another project, I have worked with speech/hearing pathologists to develop equivalent forms of a test of spoken word recognition for individuals with hearing impairment. I have also co-authored successful grant applications and carried out research activities related to the development of a measurement literacy instrument. Finally, I gained teaching experience as an assistant in an advanced educational statistics course.

Intern, Summer 2010
National Center for the Improvement of Educational Assessment (Center for Assessment), Dover, NH

I worked on a project with the Pennsylvania Department of Education to develop a handbook for validating local assessment systems. State policy allowed for local entities (e.g., school districts, Intermediate Units) to develop alternatives to the statewide

Keystone end-of-course assessments. Development of the handbook required working with a state-convened committee to identify key aspects of technical quality and to balance these aspects with feasibility considerations.

Research Analyst III, 2004 – 2008
Research Analyst I, 2002 – 2004
Student Affairs Research & Assessment, Washington State University, Pullman, WA

I provided data support for enrollment management, co-curricular, and accreditation initiatives. Work involved querying large and small databases to develop reports about undergraduate applicants, enrollment yields, and academic performance (cumulative grade point average, retention, graduation). I also oversaw and coordinated surveys of recent alumni, scholarship recipients, and non-returning students. Duties related to these surveys included item writing, questionnaire design, sampling, working with on-campus printing units, preparing mailings, conducting phone interviews, compiling and analyzing response data, and writing technical reports.

Graduate Research Assistant, 2001 – 2002
Dept. of Resource Recreation & Tourism, University of Idaho, Moscow, ID

I used survey data to prepare an interpretive management plan for a USDA Forest Service Ranger Station in Idaho. I also served as a Teaching Assistant for a new course on environmental education, and was instrumental in developing the course structure.

RESEARCH INTERESTS

The unifying theme throughout my research interests is validity, specifically the inferences that test users draw from test scores. I am interested in the means through which assessment results are communicated to stakeholders and stakeholders' abilities to understand and act upon that information. I am also interested in the concept of validity and how one develops validity arguments across different assessment contexts.

PUBLICATIONS

Gotch, C. M., & French, B. F. (in press). *Elementary teachers' knowledge and self-efficacy for measurement concepts.*

Gotch, C. M., & French, B. F. (2011). The factor structure of the CIBS-II-Readiness assessment. *Journal of Psychoeducational Assessment, 29*, 249-260.

Gotch, C. M., & Hall, T. (2004). Understanding nature-related behaviors among children through a Theory of Reasoned Action approach. *Environmental Education Research, 20*, 157-177.

PRESENTATIONS

Douval-Couetil, N., Barrett, B., Hart-Wells, L., & **Gotch, C.** (2012, October) *Differentiating undergraduates from graduate student and faculty inventors*. Paper to be presented at the 2012 Frontiers in Education Conference, Seattle, WA.

French, B. F., **Gotch, C. M.**, Mantzicopoulos, P. Y., & Valdivia Vazquez, J. A. (2012, April). *A factor model for the Brigance IED-III social-emotional scale*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Gotch, C. M., & Perie, M. (2012, April). *Using validity arguments to evaluate the technical quality of local assessments*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Duval-Couetil, N. & **Gotch, C.** (2012, January). *Characteristics of students entering a multidisciplinary entrepreneurship education program.* Paper presented at the United States Association for Small Business and Entrepreneurship conference, New Orleans, LA.

Kirk, K. I., Prusick, L.M., Silberer, A.B., Eisenberg, L.S., Young, N.M, French, B.F., Giuliani, N.P., Martinez, A.S, Ganguly, D.H., **Gotch, C. M.**, Weber L., & Stentz, S. (2011, July). *The Multimodal Lexical Sentence Test for Adults: Performance of listeners with hearing loss*. Paper presented at the Conference on Implantable Auditory Prostheses 2011, Pacific Grove, CA.

Gotch, C. M., French, B. F., Kirk, K. I., Prusick, L. M., Eisenberg, L. S., Martinez, A. S., & Ganguly, D. H. (2011, July). *Deriving equivalent forms of a multimodal lexical sentence test,* Poster presented at the 13th Symposium on Cochlear Implants in Children, Chicago, IL.

Kirk, K. I., Eisenberg, L. S., French, B. F., Prusick, L. M., Martinez, A., Ganguly, D. H., & **Gotch C. M.** (2011, July). *Development of the Multimodal Lexical Sentence Test for Children (MLST-C),* Poster presented at the 13th Symposium on Cochlear Implants in Children, Chicago, IL.

Kirk, K. I., Prusick, L. M., Silberer, A. B., Eisenberg, L., Young, N. M., French, B., Giuliani, N. P., Martinez, A., Ganguly, D. H., **Gotch, C. M.**, Weber, L., & Stentz, S. (2011, July). *The Multimodal Lexical Sentence Test for Children: Performance of children with normal hearing,* Paper presented at the 13th Symposium on Cochlear Implants in Children, Chicago, IL.

Kirk, K. I., Eisenberg, L., French, B. F., Prusick, L. M., Martinez, A., Ganguly, D. H., **Gotch, C. M.**, Silberer, A. B., & Giuliani, N. P. (2011, July). *The Multimodal Lexical Sentence Test for Children: Performance of children with hearing loss,* Paper presented at the 13th Symposium on Cochlear Implants in Children, Chicago, IL.

Gotch, C. M., & French, B. F. (2011, April). *Development of and validity evidence for the Teacher Educational Measurement Literacy Scale,* Poster presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

French, B. F., & **Gotch, C. M.** (2011, April). *Elementary teachers' knowledge and self-efficacy for measurement concepts*, Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gotch, C. M., & French, B. F. (2010, May). *Sex differences in item functioning in the Comprehensive Inventory of Basic Skills-II.* Poster presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Gotch, C. M., & French, B. F. (2010, May). *The factor structure of the CIBS-II-Readiness assessment.* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Beecher, C. C., & **Gotch, C. M.** (2010, May). *The connection between early language skills and later reading ability in at-risk children.* Invited poster presentation at the Washington State University President's Summit on Early Learning, Seattle, WA.

McCracken, V. A., Hermanson, F. M., & **Gotch, C. M.** (2005, November). *Enrollment projections: The effect of policy changes*. Presentation at the annual Strategic Enrollment Management meeting of the American Association of Collegiate Registrars and Admissions Officers, Chicago, IL.

TECHNICAL REPORTS

French, B. F., **Gotch, C. M.**, & Valdivia Vasquez, J. A. (2011). *Inventory of Early Development-III—Social emotional summary report.* Washington State University, Learning and Performance Research Center.

Gotch, C. M. (2008). *Results of the survey of 2005-2006 alumni.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2008). *Spring 2008 survey of non-returning students.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2007). *Spring 2007 survey of the Fall 1997-Summer 2002 classes.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2007). *Spring 2007 survey of non-returning students.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2006). *Results of the 2003-2004 Washington State University alumni survey.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2004). *Results of the 2001-2002 Washington State University alumni survey.* Washington State University, Student Affairs Research and Assessment.

Gotch, C. M. (2003). *1997-1998 Alumni survey results (executive summary).* Washington State University, Student Affairs Research and Assessment.

PROFESSIONAL AFFILIATIONS

American Educational Research Association, Divisions D and H
National Council on Measurement in Education
American Psychological Association, Division 5

SERVICE

Committees and Governance

American Educational Research Association
- Division D Graduate Student Seminar Committee, 2010-2011
- Division D Business Meeting and Luncheon Committee, 2009-2010

National Council on Measurement in Education
- Graduate Student Issues Committee, 2009-2012 (Chair, 2010-2012)
- Task Force to Improve the Quality of the NCME Annual Meeting, 2010-2011

Washington State University
- College of Education Research Advisory Committee, 2011
- Education Graduate Organization
  - Vice President, 2010-2011
  - Co-Treasurer, 2009-2010
- Graduate and Professional Student Association, Senator, Fall 2008 Semester

Editorial Board ad-hoc reviewer

The Teacher Educator, 2011

Conference Proposal Reviewer

National Council on Measurement in Education
- Graduate Student Poster Session, 2010, 2011, 2012

Conference Session Chair

National Council on Measurement in Education
- 2012, Chair and Moderator, *Emerging issues in graduate student preparation and the work of new professionals* (Invited symposium sponsored by the Graduate Student Issues Committee)
- 2011, Chair, *Translating technical material for lay audiences* (Invited symposium sponsored by the Graduate Student Issues Committee)

AWARDS

Mentored Doctoral Student Award for the 13th Symposium on Cochlear Implants in Children, July 2011 ($800)

Washington State University Department of Educational Leadership and Counseling Psychology Travel Grant, Spring 2011 ($400) & Spring 2012 ($400)

Washington State University Graduate and Professional Student Association Travel Grant, Spring 2012 ($215)

Washington State University Graduate and Professional Student Association Registration Grant, Spring 2010 ($100), Spring 2011 ($100), & Spring 2012 ($100)

AERA Division D Measurement and Research Methodology Graduate Student Travel Award for the 2010 Annual Meeting ($1,000)

FUNDED MEASUREMENT AND EVALUATION CONSULTANCIES

Kucer, S. B. *Fluency and the processing of expository discourse: What factors predict comprehension?* Washington State University-Vancouver (2010). statistical consultant.

This project involved basic research on the effects of reading miscues on comprehension. Data were analyzed via ordinal logistic regression.