Abstract

DEVELOPMENT AND ANALYSIS OF A MEASUREMENT SCALE FOR TEACHER ASSESSMENT LITERACY

by

Catherine Ann Howell

May 2013

Director of Thesis: Dr. Scott Methe

Major Department: School Psychology

The purpose of this study was to develop and analyze an updated scale of teacher assessment literacy that measures teachers' self-ratings of skills in areas mentioned in the current literature on assessment. Items were included in the scale based on expert judgment. The participants were 193 Kindergarten through fifth grade teachers in a rural school district in the southeastern United States. All demographic surveys and scales were distributed in classrooms or at an after school staff meeting and all were collected within a three week time frame. Following collection, data were entered and analyzed. Independent t-tests and analysis of variance indicated that statistically significant differences existed in overall scale scores as a function of education level, grade level taught, and measurement courses taken. Teachers with a higher level of education had higher overall scores than teachers with less education and teachers who have taken a course in measurement had higher overall scores than teachers who have not

taken a course in measurement. Principal components analysis indicated that all items had moderate to high loadings on to a single component, which may be called “assessment literacy.”

DEVELOPMENT AND ANALYSIS OF A MEASUREMENT SCALE FOR TEACHER ASSESSMENT LITERACY

A Thesis

Presented To the Faculty of the Department of School Psychology

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Arts in Psychology

by

Catherine Ann Howell

May 2013

DEVELOPMENT AND ANALYSIS OF A MEASUREMENT SCALE FOR TEACHER ASSESSMENT LITERACY

by

Catherine Ann Howell

APPROVED BY:

DIRECTOR OF
THESIS: _____
Scott Methe, Ph.D.

COMMITTEE MEMBER: _____
Stephen Kilgus, Ph.D

COMMITTEE MEMBER: _____
Mark Bowler, Ph.D.

CHAIR OF THE DEPARTMENT
OF PSYCHOLOGY: _____
Susan L. McCammon, Ph.D

DEAN OF THE
GRADUATE SCHOOL: _____
Paul J. Gemperline, Ph.D.

ACKNOWLEDGEMENTS

First, I would like to thank my parents for providing me with the opportunity to gain this tremendous education and always supporting me in all my endeavors. I would like to thank my husband, Joshua, for supporting me while I attended graduate school and listening to countless hours of school psychology discussion. I would like to thank the director of my thesis, Dr. Scott Methe, for always being available to read my revisions and answer my questions when I needed him. I would also like to thank my committee members, Dr. Steven Kilgus and Dr. Mark Bowler, for all of the help they provided during this process. Lastly, I would like to thank the elementary school teachers who served as participants in my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

CHAPTER I: INTRODUCTION

Assessment literacy is a vital to improving both teacher instruction and student learning. The proper use of assessment has been linked to benefits in student learning, increased levels of student achievement, and improvements in teacher instruction (Mertler, 2005; Wang, Wang, & Huang, 2007; DeLuca & Klinger, 2010). Specific methods of assessment, such as formative assessment, have also been linked to student motivation and achievement (Cauley & McMillan, 2010). However, many teachers either lack specific training in assessment or do not feel adequately prepared to assess their students' performance (Plake, Impara, & Fager, 1993; Quilter & Gallini, 2000; Mertler, 2009). Although the benefits of assessment literacy are known, it appears that many teachers lack the assessment literacy required to engage in proper assessment practices (Wang, et al. 2007). In response to this need for assessment literacy, researchers continue to examine teacher assessment literacy, what it means, and how teacher training and professional development programs can be improved (Stiggins, 1991; Plake et al., 1993; Braden, Huai, White, & Elliot, 2005; Mertler, 2005; Burry-Stock & Frazier, 2008).

Studies and articles on teacher assessment literacy have been appearing in education literature since the 1990s (Stiggins, 1991). Several studies have examined assessment literacy as an outcome of teacher training or professional development programs (Stanevich, 2009; Volante & Fazio, 2007; Wang, Wang, & Huang, 2008), whereas others have examined assessment literacy as it relates to student outcomes, such as achievement (Braney, 2010; Mazzie, 2008). Assessment literacy has also been examined by comparing the assessment literacy of different teacher populations, such as preservice versus inservice teachers (Mertler, 2003; Mertler, 2005). Most importantly for the current study, assessment literacy has been examined through the development of instruments that attempt to measure teachers' knowledge, use, or skill in various

areas of assessment (Burry-Stock & Fazier, 2008; Plake, Impara & Fager, 1993). Many of these studies vary in their purpose, but they also differ in other ways. Studies of assessment literacy have used a variety of measures or response formats and have also examined many different areas of assessment.

Assessment literacy has been examined using tests, self-report measures, and open-ended questions. Specifically, some studies use instruments that measure assessment knowledge in a quiz format, such as the *Teacher Assessment Literacy Questionnaire* (Plake, Impara & Fager, 1993). Others use instruments that measure teachers' perceptions of their own assessment literacy, such as the *Assessment Practices Inventory* (Zhang, 1995) and the *Assessment Practices Inventory Revised* (Burry-Stock & Frazier, 2008). In addition, some studies do not use a developed instrument at all, but measure teacher assessment literacy by creating open-ended surveys. Volante & Fazio (2007) asked short answer questions about the main purposes of assessment and teacher preferences, for example, to examine changes in teacher confidence levels in assessment after completing a training program.

Assessment literacy has clearly been measured in a variety of ways, which is necessary in the development of a construct, but a problem remains in the way that the construct is defined. While a common practice has been to use a list of teacher assessment standards as a blueprint (*i.e. The Standards for Teacher Competence in Educational Assessment of Students*), as a whole, studies of assessment literacy are inconsistent in the way that the construct is defined. Examination of instruments that have been empirically validated shows that, even after considering that the definition of assessment literacy is inconsistent in the literature, assessment literacy items are constantly changing to incorporate different aspects of assessment. For example, the *Assessment Practices Inventory* uses items that directly relate to *The Standards for*

*Teacher Competence in Educational Assessment of Students (*American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990), whereas the *Assessment Practices Inventory Revised* added items that were meant to be more consistent with current literature in assessment, such as items relating to formative assessment (Burry-Stock & Frazier, 2008).

The Standards for Teacher Competence in Educational Assessment of Students contains important ideas that are still very relevant to assessment literacy. In addition to this list, many studies and lists of competencies that mention additional assessment skills have appeared in more recent literature (McMillan, 2000; Stiggins, 2009; DeLuca & Klinger, 2010; Brookhart, 2011). For example, the use of formative assessment is a critical skill that appears in current literature and should be reflected more often in scales of teacher assessment literacy. The *Assessment Practices Inventory Revised,* a self-report measure of assessment literacy*,* reflects this change in its items. However, classroom assessment has also expanded to include aspects of Response to Intervention (RTI), such as progress monitoring (National Center, 2010). Not only is RTI important for legal purposes (e.g., placement decisions of children), but knowing how to assess students in an RTI framework is important for ensuring that students are receiving the appropriate level of instruction that allows them to be the most successful (Gresham, Reschly, & Shinn, 2010). Therefore, this change also needs to be considered when examining teacher assessment literacy. The purpose of this study is to develop a self-report rating scale of teacher assessment literacy that examines teachers' self- perceived skills in several areas of assessment literacy mentioned in the current literature on assessment.

Statement of the Problem and Rationale for Current Study

Research indicates that teachers spend a vast amount of time in assessment-related activities and that these activities positively affect the quality of instruction and student achievement when assessment is sound and results are meaningful and accurate (Schaffer, 1993; Black & Wiliam 1998; Mertler, 2005). Measuring teacher assessment literacy is an attempt to identify levels of knowledge and skill that are required to perform sound assessment practices. Only two empirically validated assessment instruments are used in multiple studies (the *Teacher Assessment Literacy Questionnaire/*CALI and the API/API_R). The API_R reports adequate reliability and validity through internal consistency, factor analysis, and comparison to the original API. However, the *Teacher Assessment Literacy Questionnaire/*CALI only reports internal consistency coefficients, and these data suggest that the results obtained using this instrument may not be generalizable to a larger population (Mertler, 2005). No other forms of reliability (e.g., alternate forms or test-retest) or validity (e.g., construct or criterion-related) are reported. Furthermore, many studies choose to create their own measures of assessment literacy, such as open-ended surveys, (Volante & Fazio, 2007; Cizek, Fitzgerald, & Rachor, 1995), and do not examine the psychometric properties of their measures, which creates problems similar to those of the CALI. Because not all studies have focused on the same components of assessment literacy, there is little consistency in the way that assessment literacy is measured. For example, some studies use the *Standards* as a guideline, while others choose specific areas of assessment that are to be the focus of the study, which suggests that the construct of assessment literacy varies in the way that it is conceptualized (Mertler, 2005; Volante & Fazio, 2007; Burry-Stock & Frazier, 2008).

The $API_{R,}$ the second empirically validated instrument, has strong reliability and validity. It also includes aspects of assessment that are mentioned in recent literature on assessment literacy (e.g., formative assessment), which suggests that it is a good measure of assessment literacy. However, assessment needs are constantly changing and expanding in the classroom as the framework for assessment changes (e.g., from discrepancy models to RTI models). Because the areas of assessment literacy are expanding, more current and updated teacher assessment literacy instruments are needed that give more consideration to formative assessment and assessment related to RTI. In addition, because there have been few empirically validated instruments that measure teacher assessment literacy, the construct itself could benefit from further psychometric development in the areas of reliability and validity. Appendix A summarizes the domains of assessment literacy that the *Scale of Teacher Assessment Practices* (STAP), the scale used in the current study, aims to include. Because this scale aims to measure several areas of assessment literacy, the structure of the STAP will be examined, as in similar scale development studies (Burry-Stock & Frazier, 2008).

Research Questions and Hypotheses

Because this study includes the use of a newly developed instrument, the psychometric properties of the STAP need to be examined. Both internal consistency reliability and criterion-related validity will be calculated. To examine validity, an empirically validated scale that purports to measure the same construct, the $API_{R,}$ will be used as a comparison measure. In addition, as in previous scale development studies (Zhang & Burry-Stock, 1994 & Burry-Stock & Frazier, 2008), the structure of the STAP will be examined to determine if individual items converge to form distinct components of assessment literacy. Hypotheses for the psychometric properties of the STAP are as follows:

Hypothesis 1: The items on the STAP were developed with five areas of assessment literacy in mind that were considered based on areas that have been examined in other assessment literacy studies (Mertler, 2005; Volante & Fazio, 2007; Burry-Stock & Frazier, 2008). Items were developed based on the following five areas: selection and development of assessment methods, administering, scoring, and interpreting results, using results to inform day-to-day decisions, communication of results to others, and ethical use of assessment. Therefore, results of a principal components analysis (PCA) should suggest five different components of assessment literacy.

Hypothesis 2: Scores generated by the STAP will demonstrate adequate internal consistency.

Hypothesis 3: Total score on the STAP will have a strong correlation with scores on selected items of the API_R.

Demographic variables are examined in many studies of teacher assessment literacy (Plake et al., 1993; Zhang, 1995; Zhang & Burry-Stock, 2003; Burry-Stock & Frazier, 2008). Differences in assessment literacy have been found as a function of whether or not a teacher has taken a course in measurement, for example. Other demographic variables, such as level of education, are not always examined. Therefore, this study will examine whether scores on the *Scale of Teacher Assessment Practices* (STAP) vary as a function of several demographic characteristics.

Hypothesis 4: Total score on the STAP will vary by some demographic characteristics. Specifically, score should vary by years of teaching experience, education level, and courses in measurement.

CHAPTER II: LITERATURE REVIEW

Assessment Literacy

How has assessment literacy been defined and operationalized? Assessment literacy was originally operationalized by Stiggins (1991) in terms of the characteristics of an "assessment literate" person. For example, Stiggins (1991) states that assessment literates know what it takes to produce high-quality achievement data for different forms of tests and are confident enough to ask technical questions about complicated data. Stiggins did not provide a clear definition of assessment literacy, and he stated that assessment literacy is a multidimensional concept; its meaning varies as the context changes. Later, Stiggins (1995) expanded his list to include knowing what is being assessed, why it is being assessed, how best to assess achievement of interest, what can go wrong, and how to prevent problems. As researchers began to examine the construct over the next decade, a more inclusive definition of assessment literacy was developed, which states that assessment literacy is "the possession of knowledge about the basic principles of sound assessment practice, including terminology, the development and use of assessment methodologies and techniques, familiarity with standards of quality in assessment... and familiarity with alternatives to traditional measurements of learning" (Paterno, 2001). Despite the evolution of the concept, there are key ideas that are present in all definitions of assessment literacy. Being assessment literate means having a variety of knowledge and skills that are used toward the purpose of assessment, and teachers with these skills often apply them to two forms of assessment: summative and formative. Summative assessment refers to using assessment procedures (e.g., end-of-grade tests) to make decisions of accountability or the effectiveness of previous instruction, while formative assessment consists of day-to-day activities used in the classroom that allow a teacher to adjust instruction (Popham, 2009). Examples of formative

assessment include curriculum-based measurement, everyday assignments, asking questions, and text-embedded materials (Schafer, 1993). It is has been estimated that teachers spend anywhere from one third to one half of their professional time engaged in assessment-related activities (Stiggins, 2002).

Teacher training in assessment. Despite the amount of time that teachers spend on assessment-related activities, the majority of teacher training programs do not require specific coursework in assessment, such as courses in measurement, for teacher certification (Schafer, 1993; Campbell, Murphy, & Holt, 2002; Mertler, 2005). Teachers often struggle with quantitative aspects of measurement that may be covered in these courses, such as how to interpret the statistics often reported with standardized test results. In 1993, Schaffer estimated that only about half of working teachers were likely to have had a course in assessment. Years later, Stiggins (1999) stated that only 25 of the 50 states required that teachers meet assessment competence standards and/or complete assessment coursework. Stiggins noted that the development of standards of professional competence in assessment has put pressure on programs to prepare preservice teachers in the area of assessment literacy, which is a challenge they have not met in the past (Stiggins, 1999). Furthermore, a more current study by Volante and Fazio (2007) showed that programs that do provide specific coursework in assessment and evaluation cannot assume that their students are graduating with an acceptable level of assessment literacy. These researchers examined the assessment literacy of teacher candidates during their four years of a teacher training program. Volante and Fazio (2007) found that after preservice candidates completed coursework in observation techniques, formative and summative evaluation, and documentation, their confidence in assessment practices did not change significantly.

Considering the amount of time teachers are estimated to spend on assessment-related activities in the classroom, there appears to be a mismatch between teacher training and their needs. The good news is that if a teacher does not complete training coursework in assessment, there are other opportunities for teachers to build assessment literacy, such as workshops and consultation with school psychologists, which allow teachers to build their competence (Braden et al., 2005). In a study by Stanevich (2009), researchers investigated the effects of an assessment literacy workshop on teachers' perceptions of assessment. The researchers found that teachers' perceptions were positively affected, and they gained a greater sense of confidence in their knowledge of assessment and how it is used in the classroom. However, this particular study only examined teachers' thoughts and did not examine student data to confirm teacher perceptions, so it is unclear whether teachers' assessment literacy had an effect on student learning. Braden, Huai, White, and Elliot (2005) also discussed barriers to continuing professional development (CPD) and aspects of CPD that increase it's effectiveness, such as nonstandard options that engage teachers and allow them to apply what they learn. Furthermore, these researchers found that participation in a specific program ("Assessing One and All") was associated with large changes in assessment knowledge and application.

Is assessment literacy essential to teaching? According to Brookhart (1998), classroom assessment is an extremely important teaching function that contributes to many other aspects of teaching, such as instruction and classroom management. High-quality assessment is important for attaining higher levels of student achievement (Stiggins, 1995; Mertler, 2005), and most schools are required to participate in some form of large-scale assessment to evaluate student learning and achievement. So, it would appear that proficiency with assessment, or assessment literacy, is a requisite skill for high-quality teaching (Volante & Fazio, 2007). Research also

indicates that when specific types of assessment, such as formative assessment, are used on a regular basis, it has positive effects on student achievement (Cauley & McMillan, 2010). Formative assessment techniques focus on teaching and evaluating for student understanding of the material, rather than just memorization and recall, which is often the focus of summative assessment (William, Lee, Harrison, & Black, 2004). Therefore, it is important that teachers use these strategies as frequently as possible. However, it is not enough to simply use a strategy. Teachers need assessment literacy because ethical issues and consequences can arise if they cannot use and interpret assessment procedures accurately (Pope, Green, Johnson, & Mitchell, 2009; Popham, 2009). For example, when a teacher assesses a student, the teacher must know what method (e.g., what specific test will yield the most accurate and relevant information) is the most appropriate for that student. If the test is not valid for the purpose of its use, teachers may make inappropriate decisions that adversely affect a student's performance (Popham, 2009). Teachers who are not adequately trained in assessment may also be less sensitive to score pollution (e.g., if a student performs better on a test because they were "taught to the test"), which is a serious ethical problem if used in decision making (Pope et al., 2009). In contrast, teachers with assessment literacy are less likely to make these mistakes, and are also better at communicating results and progress to parents of students (Popham, 2009).

Standards for Teacher Competence in Educational Assessment of Students

Assessment is a critical part of the teaching process; but not all types of assessment suffice in every situation (e.g., only one multiple-choice item cannot be used to make a decision about a student). Teachers not only need to be knowledgeable and skilled in different areas of assessment, they need to use assessment methods in a way that maximizes benefits for both students and teachers. In 1987, the American Federation of Teachers (AFT), National Council on

Measurement in Education (NCME), and National Education Association (NEA) developed a set of teacher competencies in assessment under the mindset that good assessment is essential to good teaching (AFT, NCME, & NEA, 1990). In collaboration, these organizations named a committee to undertake a review of the literature to determine the current levels of teacher training in assessment, as well as teacher competence in practice. They also examined the assessment needs of students and classroom activities that required knowledge of assessment. The committee used the data collected to develop the *Standards for Teacher Competence in Educational Assessment of Students,* herein known as the *Standards* (AFT, NCME, & NEA, 1990). An abbreviated version of the seven standards follow:

Teachers should be skilled in:

1. Choosing assessment methods appropriate for instructional decisions.
2. Developing assessment methods appropriate for instructional decisions.
3. Administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.
4. Using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Developing valid pupil grading procedures that use pupil assessments.
6. Communicating assessment results to students, parents, other lay audiences, and educators.
7. Recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information. (AFT, NCME, & NEA, 1990).

Since the development of the *Standards*, other informal lists of teacher competencies in assessment have been created for large-scale assessment, formative assessment, and assessment

in specific academic areas. However, to this day, the original *Standards* are the most widely cited list of standards in studies of teacher assessment literacy. There have been few attempts to integrate both large-scale and formative assessment standards into one comprehensive list, despite the fact that teachers should be skilled in both areas (Brookhart, 2010). The *Standards* have been incorporated in textbooks and courses for education, used in studies of teacher competence, and have served as the basis for instruments to measure teacher competence in assessment (Brookhart, 2011; Mertler & Campbell, 2005; Plake, Impara & Fager, 1993).

Assessment literacy has been defined and redefined over the years, with Stiggins' (1991, 1995) definitions as some of the earliest that are still cited in studies 20 years later (Mertler, 2005; Volante & Fazio, 2007; Leighton et al., 2010). The *Standards* transform these definitions into a specific checklist of the knowledge and skills a teacher should possess. After the development of the standards, researchers aimed to develop instruments to measure the phenomenon of assessment literacy using the *Standards* as a guideline. Two instruments in particular have been empirically tested and offer reports of reliability and validity: The *Teacher Assessment Literacy Questionnaire* and the *Assessment Practices Inventory* (API). In addition, many studies use adaptations of these instruments to measure assessment literacy.

History of Assessment Literacy Measurement

Development of instruments with empirical support. In 1991, researchers developed an instrument using the *Standards* as a blueprint for a questionnaire that would measure teachers' competency levels in assessment. The researchers developed the *Teacher Assessment Literacy Questionnaire* (Plake, 1993), which consists of 35 multiple-choice questions to measure teacher knowledge in different areas. The items were intended to be realistic and meaningful to teachers' use of assessment. After initial pilot-testing and revisions, the instrument was pilot-tested on a

national sample of 555 teachers and 286 administrators from 45 states. The internal consistency of the instrument was reported as .54. Although internal consistency was not high, the authors note that the instrument is meant to be criterion-referenced (teachers should improve over time), therefore low internal consistency may be acceptable (Plake, Impara, & Fager, 1993). Overall, the study found that teachers performed best on items addressing how to administer, score, and interpret assessment results (Standard 3), and teachers performed the poorest on items in the area of communicating assessment results to relevant audiences (Standard 6 [Plake, et al., 1993]). In addition to performance in specific areas of assessment, the authors examined whether performance on the *Teacher Assessment Literacy Questionnaire* was related to certain teacher characteristics or perceptions. Specifically, the study found that teachers who had taken a course in measurement scored significantly higher than teachers who had not taken a course in measurement.

Campbell, Murphy, and Holt (2002) used the *Teacher Assessment Literacy Questionnaire* (Plake, 1993) to measure the assessment literacy of preservice teachers following their completion of coursework in assessment. Results showed that the inservice teachers used in Plake et al.'s (1993) original study performed higher overall than the preservice teachers in this study. The preservice teachers performed best on items related to choosing appropriate assessment methods (Standard 1); however, the preservice teachers performed the poorest on the same items as the inservice teachers (items related to communicating results).The internal consistency was higher in this study, with a Cronbach's alpha of .74 (Campbell et al., 2002).

A similar study used a revised version of the *Teacher Assessment Literacy Questionnaire* to examine assessment literacy in 2005. Mertler (2005) investigated the assessment literacy levels of both preservice and inservice teachers with this questionnaire, which was renamed the

Classroom Assessment Literacy Inventory (CALI). The CALI consists of 35 items similar to the original instrument, with some minor changes in wording to improve clarity. Results showed that preservice teachers, once again, scored highest on items related to choosing appropriate assessment methods (Standard 1). Inservice teachers performed similarly to the teachers in Plake et al.'s (1993) study, scoring highest on items related to administering, scoring, and interpreting results (Standard 3). Both sets of teachers performed the poorest in developing valid grading procedures (Standard 5). In addition, the inservice teachers scored significantly higher than preservice teachers on five of the seven standards (Mertler, 2005). In regard to limitations, Mertler (2005) notes that the use of an instrument with fairly low reliability makes it difficult to generalize the results of this study to a larger population, and he suggests that the CALI be rewritten before use in further studies.

Shortly after the initial development of the *Teacher Assessment Literacy Questionnaire,* another instrument was developed and examined, only rather than directly measuring knowledge in a quiz format, this instrument uses self-report ratings to measure teachers' use and skill in different areas of assessment (Zhang, 1994). Researchers of this study also used the *Standards* as a guideline for creating the rating scale items. Named the *Assessment Practices Inventory* (API), this instrument asks participants to rate various assessment practices according to how often the practice is used and how skilled the participant is in that practice. Initially, the API was pilot tested three times in order to create enough items that adequately covered the *Standards*. The final version of the API contained 67 items, and slightly different versions were created to be used with both preservice and inservice teachers. Using data from 449 preservice and inservice teachers from grades K-12, the API was examined using factor analysis and the Rasch model to determine item difficulty and whether the API contained a factor structure consistent with the

Standards. The factor structure of the API preservice scale was found to adequately match the *Standards*. However, the API inservice scale did not show seven distinct factors, and the individual *Standards* showed loadings on more than one factor. In addition to examining the structure of the API, the study also examined the relationship between subject area of the teacher, years of experience, and measurement courses and assessment literacy. Results showed that teachers with more experience and training in measurement had higher scores on the API (Zhang, 1995).

The API was used in 2003 by the same researchers in a different study to examine the practices and self-perceived assessment skills of a sample of teachers from the entire workforce of two school districts in a southeastern state. A total of 297 teachers completed the survey, and results showed that assessment practices varied across grade level and subject taught. For example, secondary teachers reported using paper-and-pencil tests more often than elementary school teachers. Also, teachers who reported having taken a course in measurement had higher self-perceived skills in several areas of assessment, such as using performance measures, standardized testing, test revision, instructional improvement, and communicating results. In addition, the authors' analysis of the factor structure of the API itself suggests that, although each construct is unique in some ways, assessment practices and self-perceived skills have a large degree of overlap (r = .71) (Zhang & Burry-Stock, 2003).

The API was examined again and revised by Burry-Stock and Frazier (2008). Using data sets from past studies as well as existing senior student teacher data for a total of 620 cases, the researchers examined the factor structure and item difficulty of the API skill-based items. They selected 25 items that seemed to best represent the factors and the construct as a whole. Fifty new items were developed based on review of current literature, and the researchers chose 25 of

the new items that they felt best represented the construct. The sample and data analysis of this study were very similar to the study of the original API. The researchers selected a sample of preservice and inservice teachers from grades K-12 and used these data to examine the factor structure of the new API, which was named the *Assessment Practices Inventory Revised* (API_R). The API_R yielded six factors, which were renamed according to the area that the authors felt was best represented by the items. One significant difference between this scale and the original API is that the new scale was found to have an entire factor that represented formative assessment practices. This study also examined the same demographic variables as the original study and found similar results: teachers who had taken a course in measurement scored significantly higher on the API_R than those who had not taken a course in measurement (Burry-Stock & Frazier, 2008).

The API_R has also been used in other studies and dissertations. For example, a study conducted by Braney (2010) examined the relationship between fourth grade teachers' assessment literacy and student reading achievement and used the API_R as a measure of assessment literacy. However, Braney notes that the developers of the instrument did not provide scoring information, so Braney coded items according to three constructs that were the focus of the study (design, use, and interpretation of assessment). This study also found that teachers with more years of experience had higher total scores on the API_R. The current study will also examine similar demographic variables and their relationship to scores on the assessment literacy scale used in this study.

Instruments with little empirical support. Other studies that use the *Standards* to measure assessment literacy do not always use instruments that measure, or attempt to measure, all of the areas mentioned in the *Standards.* For example, Quilter and Gallini (2000) examined the

relationship among teachers' assessment literacy, attitudes, and past experiences with assessment using Plake's (1993) *Teacher Assessment Literacy Questionnaire.* However, this particular study chose to use only 21 items (3 related to each standard) from the original instrument. Results of this study did not suggest a strong relationship between assessment literacy and attitudes toward assessment, but the researchers did find that secondary teachers scored higher on the assessment literacy scale than elementary school teachers (Quilter & Gallini, 2000). In a very different study of teacher assessment literacy, Wang, Wang, and Huang (2008) developed a web-based model and investigated its effectiveness at promoting teacher assessment literacy. While the focus of the study was the web-based model, to measure its effectiveness, the researchers created instruments to measure both assessment knowledge and perspectives. The instruments were based on the *Standards*, but only measured two concepts related to the study: constructing/administering tests and interpreting results (Wang et al., 2008).

Other Lists of Teacher Competencies in Assessment

While the *Standards* may be one of the most widely cited sets of teacher competencies in assessment, several other lists exist. Richard Stiggins, who is well-known for his work in both assessment literacy and formative assessment, developed a set of formative assessment competencies that shares many characteristics of the competencies mentioned in the *Standards.* He suggests that to be literate in formative assessment, teachers must ask five essential questions: Why assess? Assess what? Assess how? Communicate how? Involve students how? The ability to answer these questions should result in practices that have a clear purpose, reflect clear student learning targets, results that are communicated effectively to people with vested interest in the student, and involvement of students in the assessment process (Stiggins, 2009).

Other lists of teacher competencies represent more general sets of guidelines for teachers to follow to ensure good assessment practices. A list of competencies developed by McMillan (2000) consists of 11 statements that address what assessment is or does and what makes a good assessment. For example, assessment influences student learning and contains errors. Good assessment is valid, fair, ethical, and uses multiple methods. Another set of practices called the Code of Fair Testing Practices in Education (1988) provides guidelines in four general areas of assessment: developing and selecting tests, administering and scoring, reporting and interpreting, and informing test takers. Once again, there is overlap between these competencies, or components of assessment literacy, and the *Standards.* Also, many standards exist for assessment practices in specific areas of academics such as reading and mathematics, but a comprehensive review is beyond the scope of this study.

While many sets of teacher standards for assessment practice exist, there have been few attempts to integrate the many lists into an updated, comprehensive list that can be used to assess teacher practices in assessment. However, Brookhart (2011) is one of the few to create an updated list, similar to the *Standards*, that also takes into account features of good formative assessment. Examples included in this list are, "Teachers should understand the purposes and uses of the range of available assessment options and be skilled in using them" and "Teachers should be able to articulate their interpretations of assessment results and their reasoning about the educational decisions based on assessment results to the educational populations they serve (student and his/her family, class, school, community)" (Brookhart, 2011, p. 3).

Measuring Assessment Literacy in Other Ways

Some studies of teacher assessment literacy measure components mentioned in the lists above, but they do not identify a specific list that has been used as a blueprint for the study. In

two recent studies, researchers chose to identify only a few broad components of assessment literacy to measure. Volante and Fazio (2007), who examined teacher assessment literacy as preservice teachers progressed through an education program, measured the preservice teachers' confidence levels in knowledge of the purposes of assessment and their use of different methods. Questions were also included that addressed the teachers' thoughts on their own assessment literacy and the need for professional development, which is an important aspect to consider. Knowledge and self-perception may be related, but they are not identical (Volante & Fazio, 2007). In a different but relevant study, researchers chose to focus on grading procedures, which is often included as a competency necessary for assessment literacy (Cizek, Fitzgerald, & Rachor, 1995). Both of these studies developed a questionnaire with open-ended questions, presumably so only the components of assessment literacy that were of interest were addressed. The problem with the methods used in these studies is that no reliability and validity are reported, so it is unclear to what extent the questionnaires actually measured assessment literacy.

More recently, a study of assessment literacy was conducted in Canada that aimed to measure the assessment literacy of teacher candidates (DeLuca & Klinger, 2010). While the goal of the study was to use the collected information to inform teacher education, the researchers developed an assessment literacy questionnaire as a means of identifying the pre-service teachers' confidence levels in various areas of assessment, as well as the context in which they learned the items for which they felt the most confident. The researchers identified guides for assessment practices (including the *Standards*), but did not identify what was used to develop the specific items on their questionnaire. Forty-five items measured three domains of assessment literacy: assessment practice, theory, and philosophy. The questionnaire used a 5-point Likert scale for response categories. DeLuca and Klinger (2010) also performed a factor analysis on the

instrument, which supported the three proposed domains, as well as several constructs within each domain (e.g., reliability and validity issues as a construct within "philosophy"). Teacher confidence levels were then examined within each of the constructs. This questionnaire, like the *Teacher Assessment Literacy Questionnaire* (Plake, 1993), focused on knowledge of assessment, which is different from skill or literacy in assessment practices. Overall, results showed that confidence levels were high, with most candidates being the least confident in philosophy. These results differ from those of Volante and Fazio (2007), who found overall lower levels of confidence among teacher candidates.

Related Studies of Teacher Assessment Practices

The studies mentioned thus far focus mainly on what teachers know or how confident or skilled they are in various assessment practices. However, there are several other studies that examine teachers' beliefs about assessment (e.g., what practices they most emphasize or believe are most important). For example, McMillan (2001) conducted a study to examine secondary teachers' classroom assessment practices. Teachers in grades 6-12 were the population of interest, and participants completed a questionnaire measuring the extent to which they emphasize different assessment and grading practices in the classroom. Results suggested that secondary teachers consider a multitude of factors when grading students and use assessments that measure student understanding the most. A similar study using the same type of questionnaire examined the practices of elementary school teachers in grades 3-5 (McMillan, Myran, & Workman, 2002). The same questionnaire was used, and like the previous study, results showed that teachers consider a variety of factors when grading. The study also found that the teachers in the sample most frequently used three types of assessments: constructed response, objective, and teacher-made.

Other similar studies measure teachers' conceptions of assessment. For example, two different studies used a self-report instrument that asked teachers to rate the importance of various assessment practices. A study by Brown (2003) used a questionnaire that measures teachers' conceptions of assessment in four different areas (assessment for learning or improvement, student certification, school accountability, and assessment as irrelevant). Researchers first created, and then administered a questionnaire to teachers based on these four areas. Results showed that teachers agreed that assessment influences and improves learning and teaching, but they disagreed with the conception that assessment is irrelevant (e.g., assessment is flawed, assessment has negative consequences). This same questionnaire was used in another study that examined teacher assessment beliefs and practices (Calveric, 2010). Researchers found results very similar to the results in Brown's (2003) study. Teachers agreed the most with items related to assessment for improvement and the least with items related to assessment as irrelevant.

Assessment Literacy and School Psychology

A broad context for the importance of this study involves the practice of school psychologists. The end goal of developing an instrument is to be able to accurately measure teacher assessment literacy in order to assist in its promotion. According to the NASP *Model for Comprehensive and Integrated School Psychological Services*, school psychologists should be trained in how to select appropriate instruments, administer assessments, how to interpret different types of norm and criterion-referenced scores, and how to explain assessment results to a variety of populations (NASP, 2010). School psychologists are excellent candidates for assisting teachers in the appropriate use of assessment procedures and results through consultation, where it is important to communicate information and use consultation skills to

promote necessary change (NASP, 2010). However, in order for school psychologists to engage in these practices effectively when consulting with teachers about assessment, it is necessary to know teachers' strengths and weaknesses in assessment. Knowing what the teacher knows can have important consultative implications for school psychologists. Other studies have suggested that understanding what methods and instruments are being used in schools has direct implications for preservice and professional development programs for teachers (Madaus, Rinaldi, Bigaj, & Chafouleas, 2009). If consultants know in which areas of assessment teachers are the least skilled, they will know whom to target for consultation. Furthermore, they will know *what* areas to focus on in individual consultation or inservice training programs. Not all teachers possess the assessment literacy required or feel comfortable enough to engage in the assessment-related tasks necessary to meet the ever-growing needs of schools and students (Schafer, 1993; Plake, Impara, & Fager, 1993; Wang et al., 2007; Mertler, 2009). One way to remediate this problem is to create effective professional development or inservice training programs that address the assessment literacy needs of teachers (Braden et al., 2005). The implications of this study will bear upon these specific consultative practices.

Although consultation with teachers is an integral part of a school psychologist's job description, studies of assessment literacy are scarce in the school psychology literature. Begeny and Buchanan (2010) examined the effects of teacher experience in administering assessments. Administering assessments is one aspect of teacher assessment literacy that is mentioned frequently in the literature (AFT, NCME, & NEA, 1990; Brookhart, 2011). Begeny and Buchanan (2010) noted that teachers' judgments about student achievement is highly correlated with teachers' instructional decision making. For example, a teacher's beliefs about how well a student is performing academically may influence the materials and teaching strategies that the

teacher decides to use. The researchers examined teachers' judgments of students' literacy skills for teachers with and without assessment administration experience (specifically, experience with early literacy assessments). They found that teachers with more experience had more accurate judgments about student achievement (e.g., they were able to provide better estimates of a student's skills in reading). However, judgments were still inaccurate 40-50% of the time. Results and discussion of this study suggested that if teachers are able to make more accurate judgments, they will be able to make better instructional decisions (Begeny & Buchanan, 2010). This type of study should be expanded upon to have more meaningful implications for teachers and school psychologists. Similar studies should include more aspects of assessment and should examine the hypothesis that more experience with assessment administration links directly to better instructional decisions. Another study by Dixon, Hyson, and Mahlke (2012) used a traditional Likert scale survey to examine the assessment literacy of teachers in rural school districts, but specifically, the study's focus was on testing practices. The study examined teacher opinions on test content, frequency, student impact, and tests as instructional and evaluative tools. While testing practices are a critical part of assessment, being assessment literate also requires knowing how to interpret and use results, as well as being able to communicate results to students, parents, and other educators, which was not a focus of this study. Data collected suggested that barriers exist to "going beyond the test" such as time, resources, student motivation, and support (Dixon et al., 2012). This study mentioned consultative implications for school psychologists as well that are consistent with practices mentioned previously. Knowledge of how teachers use assessment, how confident they are, or their beliefs about barriers can be used to help school psychologists understand what kind of support teachers need and what kind of professional development is needed.

With the exception of the studies by Begeny and Buchanan (2010) and Dixon et al. (2012), studies that specifically examine teacher assessment literacy or components of assessment literacy (e.g., test administration) rarely appear in the school psychology literature. Therefore, this study also serves to extend the consultation literature in school psychology and create an instrument that school psychologists can use as an efficient evaluation of teachers' assessment literacy.

CHAPTER III: METHOD

Participants

A sample of elementary school teachers (Grades K-5) was selected from across nine elementary schools in a rural school district the southeastern United States. The selected sample consisted of 214 teachers. The schools were chosen from convenience, based upon a preexisting relationship between the researcher and the school administrators. The researcher worked in and was familiar with this site, and the sample suited the purpose of the study. Many previous studies investigating assessment literacy with the use of surveys have also used convenience sampling of preservice or inservice teachers in a district due to their geographic location (Cizek, Fitzgerald, & Rachor,1995; Mertler, 2005; Quilter & Gallini, 2000). According to Gall, Gall, and Borg (2007), convenience samples are used in over 95% of studies in the social sciences.

Teachers in grades K-5 were the population of interest because while both elementary and secondary school teachers use assessment, core skills, such as reading and math skills, are first introduced in elementary school. It is critical that teachers have assessment literacy in these grades because it is a crucial time period in a student's development of these skills. The sample included both regular and special education teachers. The sample did not include teachers who only specialize in classes that are not part of the core curriculum (e.g., physical education and visual arts). This exclusion was necessary to ensure that the sample consisted of teachers who are involved in the development of students' skills in core academic areas (e.g., reading, writing, mathematics) and to ensure that the teachers of interest have had the opportunity to apply assessment practices. Demographic information was collected on all participants and included gender, age, years of teaching, highest level of education, grade taught, and subject(s) taught. Participants were also asked to indicate what, if any, coursework they have completed in

measurement. These questions address standard demographic information that is included in many research studies, including the majority of assessment literacy studies mentioned previously. The researcher of this study was also interested in looking at any demographic variables that may influence teacher assessment literacy. Demographic questions preceded the main items of the scale.

Procedure

Initial item development. Individual items were generated to measure five domains of assessment literacy (selection and development of assessment methods, administering, scoring, and interpreting results, using results to inform day-to-day decisions, communication of results to others, and ethical use of assessment). Studies have shown that initially, the number of items created should be 50-100% larger than the final pool of items (DeVillis, 2003). Some items were inevitably discarded, so a large number of items was needed (DeVillis, 2003; Chafouleas, Briesch, Riley-Tillman, & McCoach, 2009). This also allowed the researcher to be selective about which items to include that best measure the construct of assessment literacy. Construction of the items was completed with the use of several guides in scale and survey development, including suggestions on writing effective statements (Fowler, 1995; DeVellis, 2003; Czaja & Blair, 2005). The initial item pool consisted of 64 items.

Expert judgment. Following the construction of the initial item pool, six expert judges participated in a content validation of the items. Experts in the field of assessment literacy were recruited who share assessment practices as a research interest. Experts were chosen based on their presence in the literature on assessment literacy. For example, researchers who had participated as an author in studies on assessment literacy and others who had contributed to or created a list of teacher competencies in assessment were asked to participate. Judges were

contacted via email and asked to complete the task within a four week time frame. Using Qualtrics Survey Software, the judges were provided with a statement of the purpose of the study, the researcher's definition of assessment literacy, and the hypothesized components within the construct. The judges were asked to read each item and to indicate which hypothesized component of assessment literacy they believed the item belonged to, how sure they were that the item belonged there, how relevant they believed the item was to that component (Plake et al., 1993; DeVillis, 2003; Chafouleas et al., 2009), and whether or not they believed that the item was integral to the construct of assessment literacy (please see Appendix C for the statement given to the expert judges). Nineteen experts were contacted, and six completed the Qualtrics Survey. Prior to analysis, decision rules were established and later used to decide whether an item would be included in the final scale (Plake & Impara 1993; DeVillis, 2003; Chafouleas, Briesch, Riley-Tillman, & McCoach, 2009). Decision rules for expert judgment are inconsistent in the literature, ranging from 50% to 100% agreement among judges for inclusion in the final scale. Therefore, the initial criterion for inclusion in this study was a) the item must be placed into the correct hypothesized component by five out of six judges, and b) the item must be considered integral to assessment literacy by five out of six judges. Then, additional cutoffs were used for confidence and relevance ratings of each item. Previous studies have used rating scales (e.g., 1 being not confident at all and 3 being very confident) for these two ratings, but once again, the literature is unclear on cutoffs. However, because the literature appears to suggest at least 50% agreement or higher, the researcher chose to use higher criteria to attempt to chose only items that best measured the construct. For this study, the researcher retained items that averaged a rating of very confident and very relevant or higher (Hardesty & Bearden, 2004; Wu, Chin, Chen, aLi, & Tseng, 2011). Judges were also asked to give opinions on clarity and

conciseness of items as they saw fit. Items that met all cutoff criteria were revised or eliminated based on this feedback. The final item pool consisted of 30 items. In addition, six items used from the *Assessment Practices Inventory Revised* that served as validity items. Eighty-three percent of the final pool of items had 100% judge agreement on hypothesized component, and 80% of items had 100% judge agreement on whether the item was integral to assessment literacy. The lowest average confidence rating for an item was 3.17 (out of 4) and the lowest average relevance rating for an item was 3 (out of 4). Please see Appendix D for final pool of items organized by hypothesized component.

Response categories. Participants were asked to indicate their skill level regarding various assessment practices that were addressed in the scale items. Each item was presented on a 5-point Likert-type scale with anchors that range from very low (1) to very high (5). Five response categories were used because, although research indicates that as number of response categories increases so does reliability and validity, such small differences within a large number of response categories may not reflect actual differences in the construct being measured. Therefore, four to seven categories is an optimal number (Lozano, Garcia-Cueto, & Muniz, 2008). An odd number of response choices was used, because unlike traditional Likert scales where the middle number might offer a choice of "opting out" of answering the question, the middle number on this scale represents a skill level. For example, a teacher may have "acceptable" skills if they believe they are decent at a practice, but not proficient. The response "n/a" was also included as an answer choice for teachers who do not engage in a specific assessment practice or felt that an assessment practice was not applicable to them. Numbers were included in addition to the descriptors so that measures of central tendency could be calculated in data analysis. Following the scale items, six items were added as construct validation items. With

the permission of the developers, the researcher selected six items from the *Assessment Practices Inventory Revised* that were expected to correlate with two of the hypothesized components in this study's scale (DeVellis, 2003).

Criterion measure. Six items from the *Assessment Practices Inventory Revised* **(**API_R**)** were selected and added to the end of the 30 item scale. Items were chosen from this scale because it is an empirically validated scale that examines teacher assessment practices using the same type of response method as the scale in the current study. Items are skill-based and answered using self-reported ratings, as are the items on the *Scale of Teacher Assessment Practices* (STAP). Specific items were chosen from factors on the API_R that were theoretically similar to the domains that the STAP intended to measure (DeVellis, 2003). Items were chosen from the factor "teacher assessment development and application" and "formative assessment," which are similar to the domains "selection and development of assessment methods" and "using results to inform day-to-day decisions." However, because the factors and domains do not address identical assessment issues and because this study examines other psychometric properties of the STAP as well, only six items, three from each selected factor of the $API_{R,}$ were chosen. Factor loadings for these items range from 0.535-0.819 on the intended factor (Burry-Stock & Frazier, 2008).

Administration to the development sample. Prior to administration of the scale to the sample population, the primary researcher contacted the Exceptional Children's Department director of the school district, who presented the superintendent with the STAP and the intent of the study. Following IRB approval and approval from the superintendent, the primary researcher contacted the principals of each elementary school in the district to discuss the distribution of the scale. Principals were first contacted via email, and if they did not response within three days,

the primary researcher contacted them via telephone. The primary researcher also met with some of the principals in person. The researcher explained the purpose of the scale, how it would be distributed to the participating teachers within the school, and suggested days for distribution were discussed.

With the permission of each principal, the primary researcher distributed the STAP to teachers in each participating school. The STAP was distributed in one of two ways. When possible, the scale was distributed by the primary researcher at an afternoon meeting (e.g., staff meeting) and was immediately filled out by the participants and collected. Those that could not be immediately filled out were collected the following week. When it was not possible to distribute the scales at a meeting, the primary researcher personally delivered the STAP to each teacher at his or her classroom. Regardless of the location of distribution, a general script was used to ensure that each teacher received the same information and that the researcher's interaction with each teacher was similar. Each teacher received a packet containing a cover sheet and the STAP. The cover sheet listed four statements, which the teacher was asked to read and initial upon completion of the scale (found in Appendix E). For example, one statement was "I have answered all questions honestly, and to the best of my ability." These statements were included to ensure that the scale was filled out in a standardized way. Teachers were told when the researcher would return to collect remaining scales. For teachers who did not complete the STAP on the day of receipt, follow-up emails were sent the following week. One week after distribution, the primary researcher collected scales that had been completed. However, some teachers had not yet completed the scale. The researcher took note of these teachers, and returned over the next two weeks to collect the final remaining scales. Although teachers were told to hold on to the STAP until the researcher returned, some teachers opted to leave it in the front

office after completion. A total of 214 surveys were handed out, and 193 were collected for a response rate of 90%. See Appendix F for complete scale given to teachers.

A paper scale rather than a web-based scale was chosen for several reasons. Both methods have pros and cons. For example, a web-based scale is more efficient and less costly, but because it is on the computer, it may not be read by all, and it also may look different depending on what kind of computer is being used. Furthermore, studies indicate that traditional paper surveys may yield higher response rates (Kaplowitz, Hadlock, & Levine, 2004; Shih & Fan, 2008). A paper survey was also chosen because many of the assessment literacy studies mentioned previously have used paper surveys. Most have delivered surveys in classes or staff meetings, while others have mailed the survey.

Sample size. Because of a proposed principal components analysis, a large sample size was needed, and all teachers in the county who fit the criteria of the study were asked to participate. Although there is no agreed upon rule for sample size when an exploratory method such as a principal components analysis is considered, suggestions are generally for as large a sample as possible, ranging from a variable to participant ratio of 1:3 to 1:20 (DeWinter, Dodou, & Wieringa, 2009). MacCallum, Widaman, Zhang, and Hong (1999) demonstrated that in general, a larger sample size reduces sampling error, but when communalities are high (i.e., >.60) and components are well-determined (not many components or indicators of each), sample size may be below 100. When commonalities are adequate (i.e. >.50) and components are well-determined, a sample of 100-200 may be more sufficient. Because the researcher identified hypothesized components and items based on previous literature, they were estimated to be well-determined, but commonalities were not identified until data analysis. Therefore, the researcher chose to use a sample size of 200 (DeWinter, Dodou, & Wieringa, 2009; Schmitt, 2011).

Data entry. Following collection of completed scales, all demographic information was coded for data entry. All demographic information, STAP items, and API_R items were entered into SPSS for each participant. To ensure accurate data entry, the primary researcher entered all data with the help of a research assistant. One person read item answers aloud while the other entered the data. If a participant did not answer all of the items or selected "n/a" as their answer choice, data were entered as "999" and coded as "missing data." After entry of all surveys, 20% of the surveys were randomly selected and used to check the accuracy of data entry. Thirty-nine surveys were checked, and accuracy was found to be 99.996%.

Data Analyses

Eighty-one responses were missing of a total 6,948 responses, for a total of 0.01 percent missing data. Listwise deletion was used to exclude missing responses.

Descriptive statistics were calculated for demographic items. Each demographic variable was examined for potential relationships with total STAP score (as many studies examine). Descriptive statistics for individual items on the STAP were examined, as well as descriptives for API_R items used in this study. Item descriptive statistics were calculated to examine whether certain items or groups of items were rated higher or lower, on average. API_R item characteristics were compared to API_R characteristics from the original study to examine whether similar patterns occurred in each data set. Ideally, API_R item characteristics in the present study should be similar to characteristics in the original study.

A principal components analysis (PCA) was conducted for the 30 rating scale items that encompass the STAP. A confirmatory factor analysis (CFA) was not used because, while items were derived from past research, the STAP does not have a strong empirical and conceptual foundation, and all elements of the factor model were not pre-specified (Brown, 2006). PCA was

used over an exploratory factor analysis (EFA) because, while both are exploratory methods, EFA only accounts for shared variance, and this study aimed to account for all variance, including the variance that is unique to each item (Pett, Lackey, & Sullivan, 2003). A Kaiser-Meyer-Olin (KMO) test of sampling adequacy was conducted, as well as Bartlett's test of sphericity to determine if a principal components analysis was warranted. For a principal components analysis to be warranted, the KMO should be above 0.8, and Bartlett's test of sphericity should be statistically significant. The items were then examined using minimum average partial (MAP) analysis, eigenvalue > 1 criteria, and a scree plot to determine how many components should be extracted. Components with eigenvalues above one are considered for extraction. Five components were associated with eigenvalues above one. Examination of the scree plot showed a sharp drop at the first component, suggesting that only one component should be retained. Components were extracted using principal component analysis. No rotation was necessary due to the retention of only one component. All items loaded on to this component.

The reliability of the STAP was examined by calculating Cronbach's alpha. Cronbach's alpha was calculated for the entire set of 30 items to measure the internal consistency of the scale. Results were indicative of the extent to which items measure the same construct. Following reliability analysis, the relationship between participants' total score on the STAP and participants' total score on selected items of the API_R was examined using a Pearson product-moment correlation. Total scores represented the sum of items 1 through 30 of the STAP. The same equation was used to produce the sum of the six API_R items. As a result of the calculations, two additional variables were created for each participant (total score on the STAP and total score on API_R items). A Pearson product-moment correlation was calculated to examine if a

relationship existed between the two scores. Because the two sets of items are intended to measure the same construct, a positive relationship should exist between total scores.

Using independent t-tests and one-way ANOVAs when more than one group existed (and Levene's test was not statistically significant), the relationship between each demographic variable and the mean total STAP score of participants was examined. Past studies of assessment literacy (Plake et al., 1993; Zhang, 1995; Zhang & Burry-Stock, 2003; Burry-Stock & Frazier, 2008) have examined the relationship between assessment literacy scores and demographic characteristics. Some have found that assessment literacy scores do vary by some characteristics, such as number of measurement courses taken. Results of this study were looked at in terms of whether or not assessment literacy varied as a function of age, gender, number of years teaching, grade taught, education level, and number of courses taken in measurement. Also, results were looked at in terms of whether the relationships found in other studies were also found in the current study. When Levene's test was statistically significant, Brown's Forsythe and Games-Howell post hoc tests were used instead of a one-way ANOVA. Statistically significant differences were identified based on a significance level of less than 0.05. Where statistically significant differences were found between more than two groups, a Tukey's honestly significant difference (HSD) post hoc comparison test was used to identify which groups were significantly different.

CHAPTER IV: RESULTS

Participant Characteristics

One hundred and ninety-three teachers completed the STAP, including eight men (4.1%) and 184 women (95.3%); one participant did not disclose their sex. The mean age of teachers who completed the scale was 39.04 and the mean number of years teaching was 12.46. The most frequent level of education was a bachelor's degree (63.7%). These characteristics are similar to teacher characteristics reported in the U.S. Census 2007-2008 Schools and Staffing Survey. The grade level taught was evenly distributed, with slightly fewer special education teachers than any other grade (10.4%). More than half of participants had not taken a course in measurement (56.5%). Demographic information is presented in Tables 1 and 2.

Item Characteristics

Means, standard deviations, and item-to-total correlations for each item on the STAP are presented in Table 3. On average, teachers in the sample rated themselves the highest in "adhering to the bounds of confidentiality regarding assessment results." Teachers in the sample rated themselves the lowest in "sampling from the domain defined by learning goals to write assessment items."

Means and standard deviations of each API_R item were also examined. See Table 4. Comparison of descriptive statistics from this study and descriptive characteristics from the original API_R (Burry-Stock & Frazier, 2008) show that current means are slightly higher for nearly every item, and current standard deviations are slightly lower for every item. Item-to-total correlations were also calculated to determine if these correlations are similar to correlations found in the original study. The item-to-total correlations found in this study were also higher than correlations found in the original study.

Table 1.

Demographic Characteristics of Participants (N = 193)

Characteristic	*n*	*%*
Gender		
Male	8	4.14
Female	184	95.34
Highest level of education		
Bachelor's degree	123	63.73
Master's degree	66	34.20
Ph.D	1	0.52
Grade level taught		
Kindergarten	33	17.1
1st grade	29	15.0
2nd grade	28	14.5
3rd grade	27	14.0
4th grade	27	14.0
5th grade	26	13.5
Special education	20	10.4
Number of courses taken in measurement		
None	109	56.5
1-2	63	32.6
More than 2	16	8.3

Note: Totals of n are not 193 for every characteristic because of missing data. Totals of percentages are not 100 for every characteristic because of missing data.

Table 2.

Other Participant Characteristics (N = 193)

Characteristic	*M*	*SD*
Age (n = 184)	39.04	10.75
Years of teaching (n = 191)	12.46	8.99

Table 3.

Means and Standard Deviations of Individual Items

Item	*M*	*SD*	Item-to-total correlation
1. Explaining assessment results clearly to parents	3.92	0.75	0.64
2. Seeking assistance when I am unsure how to score an item	4.32	0.72	0.53
3. Choosing an assessment method for a specific purpose, relating to an individual student	3.84	0.83	0.75
4. Using the results of formative assessment to adjust the content of my lessons	4.14	0.72	0.70
5. Adhering to the bounds of confidentiality regarding assessment results	4.41	0.78	0.56
6. Using assessment results to appropriately group students for instruction	4.14	0.75	0.64
7. Selecting appropriate methods for reporting results to others, in addition to grades	3.81	0.73	0.73
8. Explaining results to other educators for the purpose of assisting with placement decisions	3.80	0.84	0.68
9. Using results of summative assessments to adjust future lesson plans	4.05	0.70	0.69
10. Knowledge of the consequences of unethical use of assessment	4.14	0.94	0.62
11. Selecting multiple methods of assessment (e.g., tests, observations)	4.11	0.72	0.76

12. Recognizing inappropriate use of assessment	3.83	0.80	0.65
13. Interpreting summary scores reported with standardized test results (e.g., mean, percentile rank)	3.66	0.81	0.67
14. Administering progress monitoring assessments	4.05	0.82	0.45
15. Creating assessments that accommodate the needs of a variety of students	3.68	0.86	0.70
16. Explaining to parents how assessment results are used to make decisions about their children	3.99	0.82	0.68
17. Determining if an assessment is aligned with required standards (e.g., state or district curriculum goals)	3.82	0.87	0.68
18. Knowledge of which externally produced assessments are current and available	3.37	0.90	0.69
19. Administering standardized assessments (e.g., standardized achievement tests)	4.19	0.82	0.51
20. Recognizing when assessment results are being used inappropriately by others	3.60	0.89	0.62
21. Communicating the results of assessments to students in a way that they can understand	3.94	0.73	0.72
22. Using assessment information to develop an instructional plan for a student	4.00	0.79	0.78
23. Using progress monitoring results to adjust instruction	4.12	0.80	0.54

24. Using assessment results to identify students with similar needs	4.10	0.69	0.61
25. Interpreting criterion-referenced scores	3.29	0.94	0.69
26. Understanding why standardized administration is necessary to interpret results of standardized tests	3.87	0.85	0.61
27. Developing assessments with different formats (e.g., multiple-choice, fill-in-blank, short answer)	3.98	0.84	0.65
28. Identifying my own legal responsibilities in regard to assessment	3.89	0.96	0.58
29. Sampling from the domain defined by learning goals to write assessment items	3.25	0.97	0.58
30. Explaining to students how assessment results will be used to assign grades	3.82	0.82	0.64

Table 4.

Comparison of Original API_R Item Characteristics and Current API_R Item Characteristics

API_R item	Original *M**	Current *M*	Original *SD**	Current *SD*	Original Item-to-total correlation*	Current Item-to-total correlation
1. Writing fill-in-the-blank/short answer questions	3.94	3.73	1.06	0.93	.0.62	0.75
2. Using assessment results when developing lesson plans	3.65	3.97	1.00	0.76	0.60	0.81
3. Revising a test based on item analysis	3.49	3.51	1.04	0.95	0.61	0.75
4. Using assessments, such as classwork, to enhance my instructional delivery	3.85	4.16	0.83	0.70	0.64	0.84
5. Using assessment results to improve teaching and learning	3.91	4.20	1.00	0.67	0.60	0.84
6. Developing assessments based on clearly defined course objectives	3.90	4.00	0.90	0.77	0.71	0.83

Note: Original descriptive statistics have been rounded to two decimal places.
*Burry-Stock & Frazier, 2008

Research Question 1

What is the component structure of the STAP? Prior to structural analysis, KMO test of sampling adequacy was conducted to evaluate the partial correlations among items. The KMO should be above 0.8 for a satisfactory analysis and the KMO showed the sampling adequacy to be .943. Bartlett's test of sphericity was conducted to ensure that the correlation matrix was not an identity matrix, and it was significant, suggesting the matrix is not an identity matrix ($|^2$ (435) = 3442.59, p = .000). These results indicated that a PCA was warranted.

All participant ratings on the 30 items of the STAP were submitted to a PCA. Communalities, which are the proportion of item variance explained by the extracted components, after extraction, indicated that nearly all items had a communality of greater than 0.5. As mentioned, initial eigenvalues revealed five components with eigenvalues greater than one. However, there was one component with a much higher eigenvalue of 13.41. The relative magnitude of the eigenvalues was examined using a scree plot (shown in Figure 1). A sharp drop at the first component suggested that only one component should be retained. This component was extracted, and because only a single component was extracted, no additional rotations were completed. Component loadings were also examined to determine what items loaded highly on the single component. The results of the PCA are shown in Table 5.

The single component accounted for 44.7% of the variance in the items, and 29 of the 30 items had loadings above 0.5. The hypothesis that five specific components of assessment literacy are represented by the scale was not necessarily supported, since all items loaded highest on one component. However, all items had moderate to high loadings on this single component, which suggests that it may represent general assessment literacy.

Items with high component loadings. Certain items of assessment literacy loaded higher on to the single component than other items. There were 13 items with loadings above 0.7, which indicates these items may strongly represent an "assessment literacy" component. These items could be further investigated in a brief version of the STAP.

Table 5.

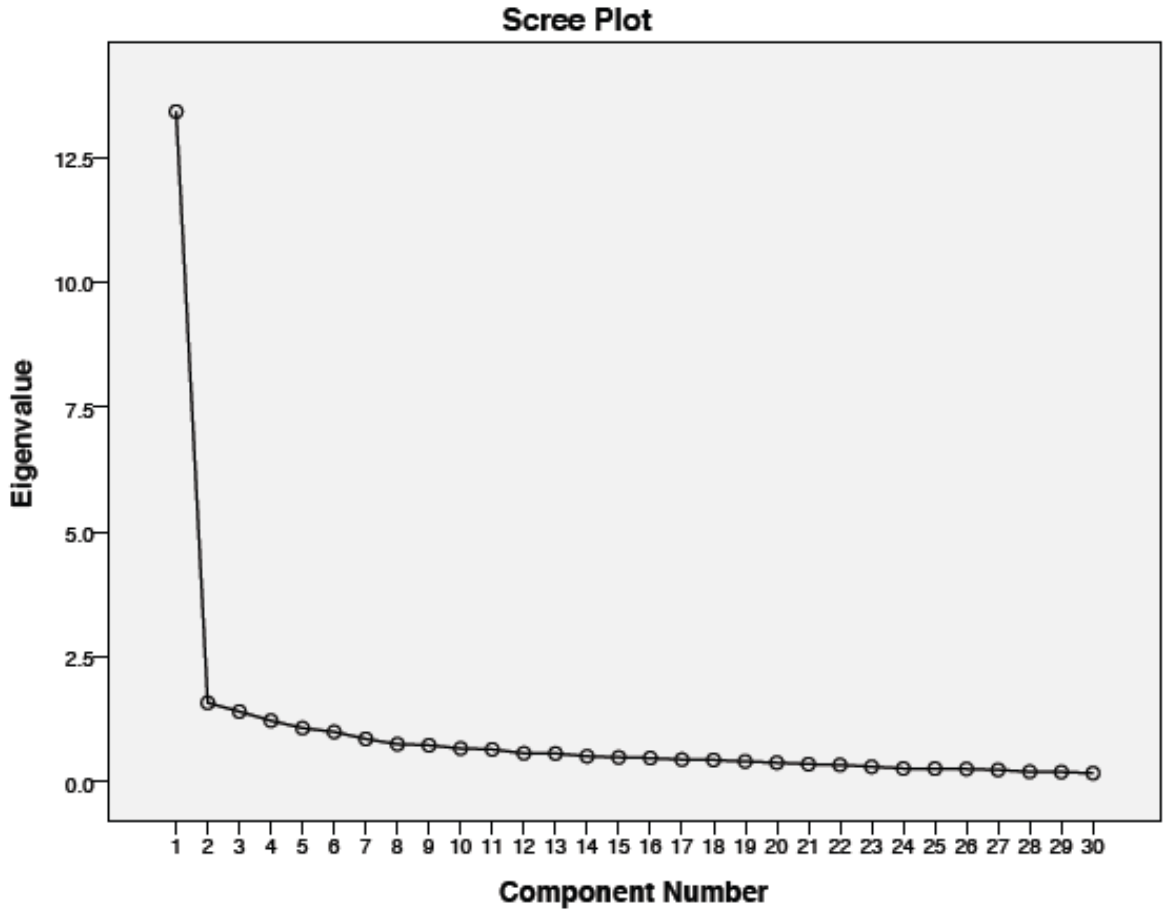
Component Loadings From Principal Component Analysis

Item	Component Loading
	1
22. Using assessment information to develop an instructional plan for a student	0.80
3. Choosing an assessment method for a specific purpose, relating to an individual student	0.77
4. Using the results of formative assessment to adjust the content of my lessons	0.75
11. Selecting multiple methods of assessment (e.g., tests, observations)	0.74
21. Communicating the results of assessments to students in a way that they can understand	0.74
16. Explaining to parents how assessment results are used to make decisions about their children	0.74
7. Selecting appropriate methods for reporting results to others, in addition to grades	0.73
9. Using results of summative assessments to adjust future lesson plans	0.73
6. Using assessment results to appropriately group students for instruction	0.72
1. Explaining assessment results clearly to parents	0.72
15. Creating assessments that accommodate the needs of a variety of students	0.71
12. Recognizing inappropriate use of assessment	0.71
8. Explaining results to other educators for the purpose of assisting with placement decisions	0.70
17. Determining if an assessment is aligned with required standards (e.g., state or district curriculum goals)	0.69
25. Interpreting criterion-referenced scores	0.68
13. Interpreting summary scores reported with standardized test results (e.g., mean, percentile rank)	0.66

	Component Loading
18. Knowledge of which externally produced assessments are current and available	0.66
24. Using assessment results to identify students with similar needs	0.66
27. Developing assessments with different formats (e.g., multiple-choice, fill-in-blank, short answer)	0.65
26. Understanding why standardized administration is necessary to interpret results of standardized tests	0.64
30. Explaining to students how assessment results will be used to assign grades	0.64
20. Recognizing when assessment results are being used inappropriately by others	0.62
5. Adhering to the bounds of confidentiality regarding assessment results	0.60
23. Using progress monitoring results to adjust instruction	0.60
10. Knowledge of the consequences of unethical use of assessment	0.58
28. Identifying my own legal responsibilities in regard to assessment	0.58
2. Seeking assistance when I am unsure how to score an item	0.56
29. Sampling from the domain defined by learning goals to write assessment items	0.54
19. Administering standardized assessments (e.g., standardized achievement tests)	0.50
14. Administering progress monitoring assessments	0.52

Figure 1.

Scree Plot From Principal Component Analysis



Research Question 2

What is the internal consistency reliability of the STAP? Cronbach's alpha was calculated for all 30 items was .96 (M = 117.46, SD = 16.42), indicating that the STAP has strong internal consistency reliability. Cronbach's alpha was also calculated for the 13 items previously mentioned with component loadings of 0.70 or higher, and was found to be .94.

Research Question 3

What is the criterion-related validity of the STAP? Does total score on the STAP correlate with scores on selected items of the API_R*?* When distribution of the summed STAP scores was examined, skewness and kurtosis were both between -1 and 1, suggesting normality in the distribution of scores (Tabachnick & Fidell, 2007). A Pearson product-moment correlation coefficient was calculated to examine the bivariate relationship between the STAP total score and the total score on selected API_R items. Previous studies' analysis of the API_R indicated that before factor rotation, the API_R appeared to be unidimensional, with all items loading on one factor. Analysis of individual factors showed that all items taken from the API_R and used as a criterion measure had factor loadings ranging from 0.535-0.819 on the intended factor (Burry-Stock & Frazier, 2008). In addition, Cronbach's alpha was reported to be .967. These data suggest that the API_R is an adequate criterion measure. There was a statistically significant, positive correlation between total STAP score and total score on selected items of the API_R, r = .701, $p < .001$. This finding supports Hypothesis 3 that total score on the STAP correlates with scores on selected items of the API_R and indicates that the STAP has acceptable criterion-related validity as it pertains to scores on selected API_R items.

Research Question 4

Does total STAP score vary by demographic characteristics of teachers? Demographic variables were analyzed to determine if groups differed on total STAP score, which was calculated by sum score on the 30 items of the STAP. Pearson product-moment correlation coefficients were calculated for age, years of teaching, total STAP score, and total API_R score. There was no statistically significant correlation between total STAP score or API_R score and these demographic variables. Correlations are shown in Table 6.

An independent samples t-test was conducted to determine if total STAP score varied significantly by sex or education level. Levene's test for homogeneity of variances was statistically significant for education level *(p* = .002), suggesting that equal variances cannot be assumed for these groups. When equal variances are not assumed, a statistically significant difference was found between teachers with a master's level education and teachers with a bachelor's level education, $t(188) = -2.42$, $p = .016$. On average, teachers with a master's degree scored higher than teachers with a bachelor's degree (master's degree: $M = 119.12$, $SD = 18.99$; bachelor's degree: $M = 113.31$, $SD = 13.70$). Figure 2 highlights the difference between these two groups. No statistically significant difference was found between groups as a function of sex, $t(190) = 1.21$, $p = .23$.

A one-way ANOVA was conducted to determine if total STAP score varied by grade level taught. Levene's test was statistically significant *(p* = .032), so Brown's Forsythe was calculated and a statistically significant difference was found between groups, $F(6, 135.09) = 2.44$, $p = .028$. Post hoc analysis using a Games-Howell test indicated a statistically significant difference in total STAP score between Kindergarten and fourth grade teachers ($p = .037$). Fourth grade teachers' total STAP score ($M = 123.74$, $SD = 16.18$) was found to be significantly

higher than Kindergarten teachers' total STAP score (M = 110.18, SD = 17.43). Figure 3 highlights the difference between these two groups. Following a Levene's test of homogeneity of variances (p = .961), an ANOVA was also conducted for schools to determine if there were any differences in total STAP score as a function of the school at which the participant taught. No statistically significant differences were found between groups as a function of school (F (8, 184) = 1.445, p = .180).

In addition to demographic variables, differences in total STAP score and the method by which the scale was completed and returned were analyzed. Following Levene's tests that showed that p was not statistically significant, and an independent samples t-test was conducted for both variables and total STAP score. No statistically significant difference in scores was found between teachers who completed the scale on their own time and teachers who completed the scale on the spot, t(191) = .853, p = .395). No statistically significant difference in scores was found between teachers who turned in their scale directly to the researcher and teachers who turned in their scale to the front office, t(191) = .085, p = .933. This may indicate that neither the location, time, or where the scale was turned in had an impact on teacher responses.

The relationship between number of measurement courses a teacher had taken and total STAP score was also examined. Levene's test was statistically significant (p = .005), and an independent samples t-test was conducted. The t-test indicated that when equal variances are not assumed, a statistically significant difference exists between the mean total STAP score of teachers who have not taken a measurement course and those who have taken a measurement course, t(183.82) = -2.53, p = .012. On average, teachers who had taken a course in measurement scored higher than teachers who had not taken a course in measurement (course: M = 118.91, SD = 18.67; no course: M = 112.68, SD = 15.06). A one-way ANOVA was also used to examine

whether total STAP score varied specifically by the number of measurement courses taken (none, one to two, or two or more). Levene's test was statistically significant (p = .016), so Brown's Forsythe was again calculated to examine mean differences. A statistically significant difference was found between groups $F(2, 120.95) = 5.24$, $p = 0.007$. Post hoc analysis using Games-Howell indicated a statistically significant difference in total STAP score between teachers who have not taken a course in measurement and teachers who have taken more than 2 courses (p = .007). The mean total STAP score of teachers who have taken more than 2 measurement courses (M = 123.63, SD = 17.48) was found to be significantly higher than the mean total STAP score of teachers who have taken no measurement courses (M = 112.68, SD = 18.67). This difference is shown in Figure 4.

Based on results that showed a statistically significant difference in total STAP scores as a function of education level and whether or not a measurement course was taken, a Pearson chi-square test was conducted to determine if a relationship exists between education level and measurement courses. A statistically significant relationship was found between the two variables. Teachers with a master's degree were more likely to have taken a course in measurement than teachers with only a bachelor's degree, χ^2 (2, 187) = 11.94, p = .003. See Table 7 for specific percentages of teachers with a bachelor's or master's degree that have and have not taken a course in measurement.

Table 6.

Correlations of Age, Years of Teaching, Total STAP Score, and Total API_R score

	1	2	3	
Variable				
1. Age	1.00			
2. Years of teaching	.791**	1.00		
3. Total STAP score	-0.072	-0.109	1.00	
4. Total API_R score	-0.060	-0.047	0.701**	1.00

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

Table 7.

Prevalence of Teachers With a Bachelor's (n = 121) and Master's Degree (n = 65) Within Teachers Who Have and Have Not Taken a Course in Measurement

	Bachelor's Degree		Master's Degree	
Measurement Course	n	%	n	%
No	81	66.94	27	41.54
Yes	40	33.06	38	58.46

Figure 2.

Mean Total STAP Score by Education Level

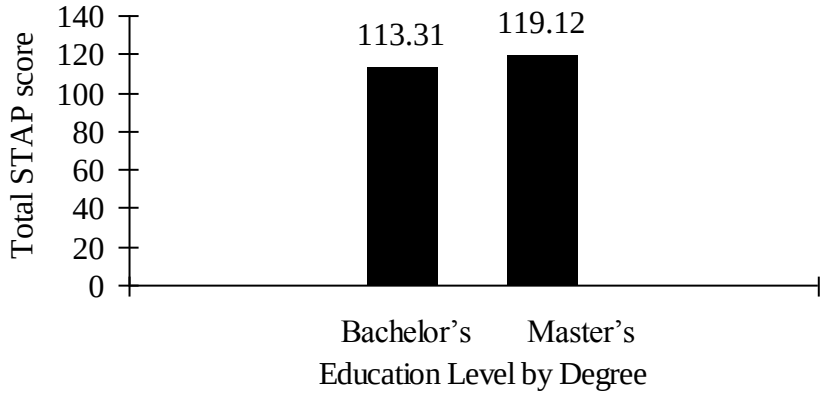


Figure 3.

Mean Total STAP Score of Kindergarten and Fourth Grade Teachers

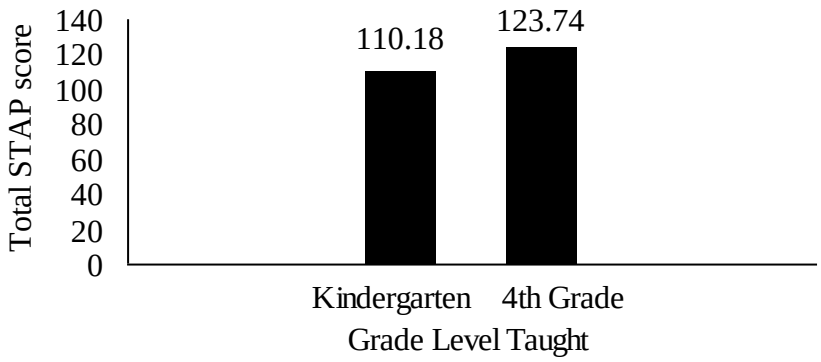
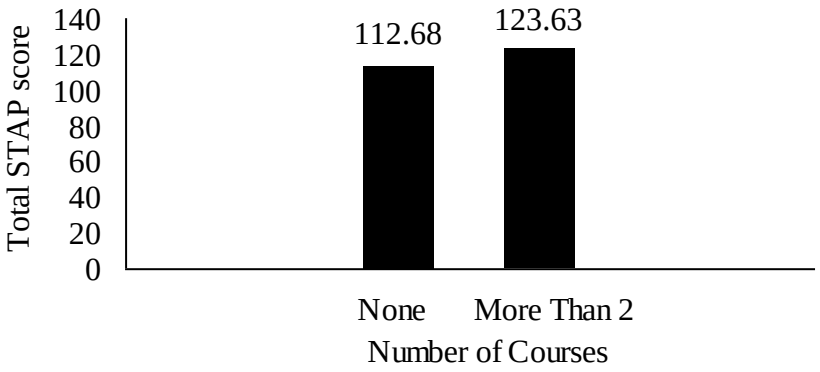


Figure 4.

Mean Total STAP Score of Teachers Who Have and Have Not Taken a Course in Measurement



CHAPTER V: DISCUSSION

Given problems or inconsistencies in other approaches, this study created a self-report measure of teacher self-perceived assessment skills. Teachers' self-perceived assessment skills, or assessment literacy, were investigated within the framework of previously established lists of teacher competencies in assessment and the current literature on teacher assessment literacy and practices. The *Scale of Teacher Assessment Practices* (STAP) was created and analyzed within current frameworks of assessment. Differences in total scale scores were examined using independent samples t-tests, ANOVAs, and Brown's Forsythe to determine how they may vary as a function of various demographic characteristics, such as age, level of education, years of teaching experience, and number of measurement courses taken. The reliability and validity of the STAP were examined through internal consistency and criterion-related validity. A factor analytical technique, principal component analysis, was applied to determine the underlying component structure of the STAP.

Results from the principal component analysis showed that all 30 items on the STAP converge on one distinct component. All items had moderate to high loadings on this component, and all loadings are above 0.5 except for one item. Although it was expected that all items should correlate since all items were meant to represent the overall construct of assessment literacy, the hypothesis that five distinct components would emerge that lined up with each domain of assessment practices addressed in the STAP was not supported. Other studies that examine similar scales, however, have found distinct factors to exist among items similar to those included in the STAP (Zhang, 1995; Burry-Stock & Frazier, 2008). For example, the two factors on the API_R that were examined contained items that were similar in content to items in two hypothesized components on the STAP (e.g., "teacher assessment development and application"

on the API_R contained similar items to "selection and development of assessment methods" on the STAP). Even though items on the STAP did not correlate into distinct components, results suggest that the scale as a whole is a strong measure of assessment literacy, which is encouraging because it allows researchers to examine and discuss assessment literacy as a unidimensional construct.

Distinct components of assessment literacy may not have been found in the STAP for several reasons. For one, there were originally ten different domains of assessment that the STAP was intended to measure. The researcher combined some areas that appeared to overlap, such as using assessment results to make decisions about students and using assessment results to enhance instruction, but questions were still created to address several different types of assessment practices. The STAP attempted to include more aspects of assessment the previous instruments, yet the number of items was reduced. So, it may be that the STAP attempted to measure too many areas within too few items. Another possible explanation may be that the sample used in this study was too narrow. Previous studies (Plake, 1993; Zhang, 1995; Mertler 2005; Burry-Stock & Frazier, 2008) examined teacher assessment literacy and practices using teachers from a broader sample including both urban and rural teachers from grades K-12, while the sample in this study included only teachers from a rural area from grades K-5. Perhaps teachers in this population were not able to distinguish items belonging to different types of practices, and it appears that participants' estimation of their skills is consistent across a range of assessment-related practices. It is also important to consider the fact that participants in this study also rated themselves higher on API_R items than did participants in Burry-Stock and Frazier's original study, and the item-to-total correlations in this study were higher than those in

the original study. This lends support to the idea that perhaps teachers in the current sample were not able to distinguish items belonging to different types of practices.

The results of the examination of the psychometric properties of the STAP demonstrated internal consistency reliability and criterion-related validity. Cronbach's alpha was .96 for all 30 items, which shows that the STAP has good internal consistency. The criterion-related validity of the STAP was determined by examining the relationship between total score on the STAP and total score on six selected items on the API_R. A coefficient of .701, $p = .000$ shows that the STAP correlates highly with the API_R and has acceptable criterion-related validity.

The results of this study lend support to the conclusion that differences in self-perceived assessment literacy differ by some, but not all, characteristics of the participants. While no statistically significant difference was found for age, sex, or years of teaching experience, significant differences were found in total STAP score as a function of grade level taught, education level, and number of measurement courses taken. Specifically, a statistically significant difference was found between Kindergarten teachers and fourth grade teachers self-reported assessment skills. Past studies of assessment literacy have generally examined the differences between either preservice and inservice teachers' scores or elementary and secondary teachers' scores. Overall, findings suggest that teachers in higher grade levels may engage in different practices and have higher self-perceived skills in assessment (Campbell et al., 2002; Zhang & Burry-Stock, 2003). In this study, fourth grade teachers had higher total scores on the STAP than Kindergarten teachers, which may be explained by the fact that in fourth grade classrooms, teachers must administer end of year tests, and therefore may have a greater concern of assessment quality throughout the year. In addition, analysis of specific answers on the STAP suggests that Kindergarten teachers simply do not engage in some assessment practices, such as

administering standardized assessments. Less experience with assessment could certainly lead to lower self-perceived skills.

The statistically significant difference found between teachers with a bachelor's degree and teachers with a master's degree lends support to the conclusion that teachers with a higher level of education may have higher levels of assessment literacy. However, this finding has not been corroborated in previous studies, as these studies have not specifically looked at the relationship between education level and assessment literacy. The finding that a statistically significant difference exists between teachers who have taken a course in measurement and those who have not taken a course in measurement supports the conclusion that teachers who have taken a course in measurement have higher levels of self-perceived assessment literacy. Many previous studies of assessment literacy also support this finding (Plake, 1993; Zhang, 1995; Zhang & Burry-Stock 2003, Burry-Stock & Frazier, 2008; Braney, 2010). However, one previous study that examined teacher confidence in assessment practices after specific training found that teacher confidence levels did *not* significantly change (Volante & Fazio, 2007). This suggests that not all assessment coursework may have the same effect on teachers' perceived skills or confidence. Assessment coursework may vary in content and rigor for a variety of reasons, such as differences in the objectives of the course or professors beliefs and opinions about assessment. It is also important to note that a statistically significant relationship exists between level of education and whether or not a course in measurement was taken. It may be that teachers with more education are more likely to take courses or have greater opportunities to take courses in measurement.

Limitations and Future Research

Although one demographic finding, differences in assessment literacy as a function of measurement coursework, has been demonstrated in many previous studies, demographic findings may have been influenced by the effect of priming, which occurs when a person becomes more sensitive to a certain stimulus as a result of a prior experience (Sintov & Prescott, 2011). Item priming effects can occur if responding to a set of items influences responses on a set of subsequent items (Sintov & Prescott, 2011). Because demographic items were placed at the beginning of the scale, they may have affected how participants answered the items. For example, a person who fills out the scale and remembers that they took several courses in measurement may be inclined to answer items more highly. Priming may result in inflated scores simply due to the fact that participants were thinking about their educational level or experience with measurement prior to answering the items. Previous research has yielded mixed results in identifying whether item priming occurs. For example, while Sintov and Prescott found weak evidence of order effects, another study that examined the effects of ethnicity priming found that increased self-awareness of ethnicity did influence responses (Forehand & Deshpande, 2001). Future research on assessment literacy should randomly alter the placement of demographic questions to control for priming as a possible confound variable in self-report measures.

Another limitation may be the sample used in this study. The sample only consisted of elementary school teachers in a rural school district, the majority of whom were women. Therefore, the results of the study may not be generalizable to the entire population of teachers in the United States. Although the sample was not necessarily small and is considered adequate for a factor analysis or a principal component analysis (De Winter et al., 2009), the teachers included in the study all teach in the same county, which is in a rural area. In addition, the assessment

literacy of only elementary school teachers was examined, and previous research has suggested that secondary teachers may engage in different practices or have a different level of assessment literacy than other types of teachers (Mertler, 2005). Also, there were very few men in this study's sample. Although a statistically significant difference in total STAP score was not found between men and women, this may be because the small number of men in the sample made a difference difficult to detect. Therefore, future research with the STAP should broaden the sample to include secondary teachers and teachers in urban school districts. Studies should examine whether differences in assessment literacy exist between elementary and secondary school teachers and rural and urban school teachers. Although it may be difficult to control, studies should also strive for a larger percentage of male participants. This research would be useful in confirming whether any differences in assessment literacy between these populations fits with previous research on assessment literacy.

The type of response method (how participants were asked to answer the items) used in this study may be a limitation to the accuracy and variability of responses in this study. Self-report scales may be subject to bias because of factors such as social desirability. Socially desirable responding occurs when a respondent answers items in a way that shows the respondent in a favorable light, such as underreporting negative behaviors and over reporting positive, or desirable behaviors (Sintov & Prescott, 2011). For example, even though teachers were assured that individual responses would not be shared with anyone other than those directly involved in the study, teachers may have rated themselves differently to appear more assessment literate to the researcher. In this case, teachers may have given themselves higher ratings on items, since a higher rating would indicate a higher level of assessment literacy. If teachers consistently rated themselves positively on most items, these ratings may have reduced variance

across items, decreasing the likelihood of differentiating components on the STAP. Future research should use multiple methods of data collection, such as direct observation of teachers in the classroom, teacher interviews, or visual analysis of tests or progress monitoring graphs to validate teacher self-reports (Zhang & Burry-Stock, 2003).

Another possible limitation to the content validity of the STAP may be the number of items included. The researcher began with 64 items, and 30 items were chosen based on expert judgment. The final pool of items was intended to be around 30 to increase the likelihood that teachers would complete the scale, and 64 items were included in the initial item pool so that the researcher could be selective in choosing the items that best measured the construct of assessment literacy. However, perhaps in order for a greater number of components to be extracted in a principal component analysis, a larger number of items should have been included in the final scale. In the final scale, each hypothesized component contained between five to seven items, which may not have been enough items to distinguish that domain as a distinct component. Future research should consider using a larger number of items per domain in order to better represent each area of assessment literacy.

Implications

This study has important implications for the professional development of teachers, the role of school psychologists, and the measurement of teachers' skills in assessment. Teacher assessment literacy is strongly linked to student learning and achievement, and when well-designed assessment is used as intended, it has positive educational significance for students, as well as teachers (Braden et al. 2005; Wang et al., 2008). Despite the positive impact that teacher assessment literacy has in the classroom, many teacher education programs do not require a course in assessment, and many states do not require these kinds of courses for licensure

(Schaffer, 1993; Cizek et al, 1995; Braden et al. 2005). The results of this study, in corroboration with the results of most previous studies (Plake, 1993; Zhang, 1995; Zhang & Burry-Stock 2003, Burry-Stock & Frazier, 2008; Braney, 2010), suggest that teachers who have had a course in measurement have higher self-perceived assessment literacy than teachers who have not had a course in measurement.

These results have important implications for the training and professional development of teachers. One solution is to increase the availability of measurement courses for preservice teachers. If teachers have better training in assessment, this may increase their efficiency and accuracy with assessment in the classroom, which may promote student achievement (Mertler, 2005; Wang, Wang, & Huang, 2007; DeLuca & Klinger, 2010). In addition, better teacher training may also affect how school psychologists use their time. Rather than promoting teacher assessment literacy, school psychologists can spend more time consulting with teachers about specific students and interventions. Improving the lives of students is the ultimate goal of consultation, and if teachers are already equipped to engage in high quality assessment, school psychologists can more directly focus on specific student needs.

Even if teacher training incorporates more measurement training, there will still be many inservice teachers who have not had training in measurement. In this case, more professional development opportunities should be available. The target of professional development for teachers should take into consideration grade level and subject area to create programs and activities that are the most applicable to attendees. Professional development should also include education in several areas of assessment, and consider that assessment can be used for very different purposes. In addition, because formative assessment and RTI have become increasingly important in the schools, teachers need to be educated in how these changes have expanded the

need for assessment. The best way to determine the areas of assessment that professional development should focus on is to use responses obtained from the teachers themselves. For example, studies using the CALI (Mertler, 2003; Mertler 2005) have consistently found that teachers in these samples have the most difficulty with developing valid grading procedures. While individual components were not found in the STAP, specific items can be examined to determine the types of practices that teachers feel they are the least skilled in. Training and professional development programs should consider the results of these studies when developing a curriculum.

The idea of professional development for teachers also has important implications for school psychologists. As previously mentioned, school psychologists are experts in both assessment and consultation. While school psychologists use different types of assessment for different purposes, they are trained in the proper selection and use of instruments and how to appropriately interpret and use results, which is applicable to all educators, especially classroom teachers. School psychologists are also experts in consultation and are often experienced in developing and leading training workshops (NASP, 2010). Therefore, school psychologists are perfect candidates for leading training workshops for teachers that address assessment issues. Scales such as the CALI, $API_{R,}$ and STAP can serve as a "needs assessment" for teachers who attend this type of professional development, so that the school psychologist can tailor workshops directly to the needs of the teachers.

One type of assessment that both school psychologists and teachers use is curriculum-based measurement (CBM), which is an extremely useful measure that may be used as part of formative assessment and can aid in daily data-based decision making. Research has suggested that when teachers implement CBM more accurately or teachers have higher-acceptability of

CBM, students make greater gains in achievement in areas such as mathematics, for example (Allinder, 1996; Allinder & Oats, 1997). With the use of an instrument such as the STAP, school psychologists can determine which teachers need support in areas of formative assessment and provide support in the administration and use of data from CBM measures. CBM is extremely important for student achievement, because information from CBM measures can be used to group students appropriately for instruction and to determine the stage of learning that a student is in in regard to an academic skill (e.g., acquisition, fluency). It is also very sensitive to student growth (Clarke, 2009). Therefore, school psychologists should aim to support the use of measures such as CBM, especially with teachers who may have less assessment literacy in this area. If teachers are able to use CBM more frequently and accurately, they will be able to provide more appropriate instruction and interventions. Once teachers know where a student stands academically, there are several group interventions that are feasible for teachers to implement that school psychologists can assist with, such as the HELPS program (Helping Early Literacy with Practice Strategies; Begeny, 2011). Overall, assessing teacher assessment literacy can assist school psychologists in supporting teachers' administration and use assessment results for instructional decisions, which will promote student achievement.

Results of this study also have implications for the use of specific items on the scale. The scale used in this study includes items that addressed a variety of assessment issues, and all items were found to load moderately to highly on to a single component. Although items did not cluster into five distinct components similar to the domains it measures, the results of principal component and expert judgment suggest that all items do correlate with one overall construct, which may be called "assessment literacy." The results of this study strongly suggest that there are some items that load very highly and may better represent assessment literacy than items that

do not have as high loadings. Specifically, items such as "using assessment information to develop an instructional plan for a student" and "choosing an assessment method for a specific purpose, relating to an individual student" had very high component loadings, indicating that they strongly represent assessment literacy. School psychologists or administrators who are wishing to obtain a quick measure of teacher assessment literacy might select only items from this scale that are above a certain cutoff (e.g., loadings of .7 or higher). However, if items from the STAP were taken to create a brief measure, the reliability and the validity of this new measure would need to be investigated to ensure that the brief STAP still has adequate psychometric properties. Future researchers should consider creating a brief version of teacher assessment literacy, as it might be more efficient and easier to use in the schools.

Lastly, this study provides implications for the use of a new assessment literacy measure. Although the STAP did not prove to have distinct components when administered to the sample in this study, a unidimensional measure of assessment literacy may be useful in applied settings. An instrument that broadly measures teachers' skills in assessment may be useful in schools that may not have the resources to provide training in distinct areas of assessment, but want to gauge teachers' skills in assessment on the whole. It could also be useful as a criterion-referenced measure that broadly examines teachers' skills in assessment before and after training or professional development to determine if teachers believe that their skills have improved.

Overall, this study extends the literature on teacher assessment literacy and offers an updated self-report rating scale for the measure of assessment literacy. The literature has shown that although teacher assessment literacy has positive effects for both students and teachers, many teachers do not have an adequate level of assessment literacy. By first measuring and then

using scales in an attempt to improve teacher assessment literacy, student learning and achievement can also be improved.

References

Allinder, R. M. (1996). When some is not better than none: Effects of differential implementation of curriculum-based measurement. *Exceptional Children, 62*(6), 525-535.

Allinder, R. M. (1997). Effects of acceptability on teachers' implementation of curriculum-based measurement and student achievement in mathematics computation. *Remedial and Special Education, 18*(2), 113-120.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: National Council on Measurement in Education.

Barkley, R. A. (2011). *Barkley Functional Impairment Scale (BFIS)*. New York, NY: The Guilford Press.

Begeny, J. C. (2011). Effects of the helping early literacy with practice strategies (HELPS) reading fluency program when implemented at different frequencies. *School Psychology Review, (40*)1, 149-157.

Begeny, J. C. & Buchanan, H. (2010). Teachers' judgements of students' early literacy skills measured by the early literacy skills assessment: Comparisons of teachers with and without assessment administration experience. *Psychology in the Schools, 47*(8),859-868.

Black, P. & Wiliam, D (1998). Assessment and Classroom Learning. *Assessment in Education, 5* (1), 7-71.

Braden, J. P., Huai, N., White, J. L., & Elliot, S. N. (2005). Effective professional development to support inclusive large-scale assessment practices for all children. *Assessment for Effective Intervention, 30*(4), 63-71.

Braney, B. T. (2010). *An examination of fourth grade teachers' assessment literacy and its relationship to students' reading achievement.* Retrieved from ProQuest Digital Dissertations. (AAT 3434570).

Brookhart, S. M. (1998). *Teaching about grading and communicating assessment results.* Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA. (ERIC Document Reproduction Service No. 419838).

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brown, G. T. L. (2003). Teachers' conceptions of assessment. Retrieved from ProQuest Digital Dissertations. (AAT 3189277).

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

Burry-Stock, J. A., & Frazier, C. H. (2008, March). *Revision of the assessment practice inventory (APIR): A combined exploratory factor analysis and polytomous IRT approach.* Paper Presented at the American Educational Research Association, New York, NY.

Calveric, S. B. (2010). *Elementary teachers' assessment beliefs and practices*. Retrieved from ProQuest Digital Dissertations. (AAT 3443729).

Campbell, C., Murphy, J.A., & Holt, J.K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Cauley, K. M. & McMillan J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House, 83*(1), 1-6.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., & McCoach, D. B. (2009). Moving beyond the assessment of treatment acceptability: An examination of the factor structure of the usage rating profile-intervention (URP-I). *School Psychology Quarterly, 24*(1), 36-47.

Cizek, G. J., Fitzgerald, S., Shawn, M., & Rachor, R. E. (1995). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Clarke, S. (2009). Using curriculum-based measurement to improve achievement. *Principal, 88* (3), 30-33.

Czaja, R. & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.

Code of Fair Testing Practices in Education. (1988) Washington, D.C.: Joint Committee on Testing Practices.

DeLuca, C. & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 1* (4), 419-438.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Newbury Park, CA: Sage Publications, Inc.

De Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*, 147-181.

Dixon, R. J., Hyson, D. M., & Mahlke, A. G. (February, 2012). Assessment literacy and RtI: Advancing teacher development. Unpublished paper presented at the National Association of School Psychology Annual Convention, Philadelphia, PA.

Forehand, M. R. & Deshpande, R. (2001). What we see makes us who we are: Priming ethnic self-awareness and advertising response. Journal of Marketing Research, 38(3), 336-348.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage Publications, Inc.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research, an introduction,* eighth edition. Boston, MA: Pearson Education, Inc.

Gresham, F., Reschly, D., & Shinn, M. R. (2010). RTI as a driving force in educational improvement: Research, legal, and practice perspectives. In Shinn, M. R. & Walker, H. M. (Eds.), *Interventions for achievement and behavior problems in a three-tier model including RTI.* Bethesda, MD: National Association of School Psychologists.

Hardesty, D. M. & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research, 52*(2), 98-107.

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly, 68*(1), 94-101.

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, 17*(1), 7-21.

Lozano, L. M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73-79.

National Association of School Psychologists. (2010). *Model for Comprehensive and Integrated School Psychological Services.* Bethesda, MD: Author.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Madaus, J., Rinaldi, C., Bigaj, S., & Chafouleas, S. M. (2009). An examination of current assessment practices in northeastern school districts. *Assessment for Effective Intervention, 34*(2), 86-93.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, 7(8).

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

McMillan, J. H. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 9*5(4), 203-213.

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

Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, *33*(2), 76-92.

Mertler, C. A., & Campbell, C. (2005 April). Measuring teachers' knowledge and application of classroom assessment concepts: Development of the "Assessment Literacy Inventory." Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, QC.

National Center on Response to Intervention. (2010). *What is RTI?* Washington, DC: Office of Special Education Programs. (ERIC Document Reproduction Service No. ED 526 859)

Paterno, J. (2001). Measuring success: A glossary of assessment terms. Building cathedrals: Compassion for the 21st century. Retrieved from www.angelfire.comlwa2/ buildingcathedrals/measuringsuccess.html

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.

Plake, B.S. (1993). Teacher assessment literacy: Teacher's competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21-27.

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, 39.

Pope, N., Green, S. K. Johnson, R. L., & Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment. *Tecahing and Teacher Education, 25*, 778-782.

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

Quilter, S. M. & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator, 36*(2), 115-131.

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice, 32*(2), 118-125.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321.

Shih, T. & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods, 20*(3), 249-271.

Sintov, N. D. & Prescott, C. A. (2001). The influence of social desirability and item priming effects on reports or proenvironmental behavior. Ecopsychology, 3(4), 257-267.

Stanevich, C. (2009*). Building assessment literacy in teachers to promote student achievement.* Retrieved from ProQuest Digital Dissertations. (AAT 1462074).

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.

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

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-27.

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

Stiggins, R. J. (2009). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment*. New York: Routledge.

Stiggins, R. J. & DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan, 90*, 640-644.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Volante, L. & Fazio X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Candian Journal of Education*, 30(3), 749-770.

Wang, T., Wang, K., & Huang, S. (2008). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education, 51*, 448 462.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education*, 11(1), 49-65.

Worthington, R. L. & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838.

Wu, L., Chin, C., Chen, C., Lai, F., & Tseng, Y. (2011). Development and validation of the pediatric cancer coping scale. *Journal of Advanced Nursing, 67*(5), 1142-1151.

Zhang, Z., & Burry-Stock, J. A. (1994). *Assessment Practices Inventory*. Tuscaloosa, AL: The University of Alabama.

Zhang, Z. (1995). *Investigating teachers perceived assessment practices and assessment competencies on the assessment practices inventory (API).* The University of Alabama. *ProQuest Dissertations and Theses,* , 163 p. http://search.proquest.com/docview/304160100?accountid=10639.

Zhang, Z. & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323-342.

Appendix A

EAST CAROLINA UNIVERSITY
University & Medical Center Institutional Review Board Office
1L-09 Brody Medical Sciences Building· Mail Stop 682
600 Moye Boulevard · Greenville, NC 27834
Office **252-744-2914** · Fax **252-744-2284** · **www.ecu.edu/irb**

Notification of Exempt Certification

From:	Social/Behavioral IRB
To:	Catherine Cruess
CC:	Scott Methe
Date:	3/7/2012
Re:	UMCIRB 11-001373 Teacher Assessment Literacy: Development and Analysis of a Self-Report Measure

I am pleased to inform you that your research submission has been certified as exempt on 3/6/2012. This study is eligible for Exempt Certification under category #2.

It is your responsibility to ensure that this research is conducted in the manner reported in your application and/or protocol, as well as being consistent with the ethical principles of the Belmont Report and your profession.

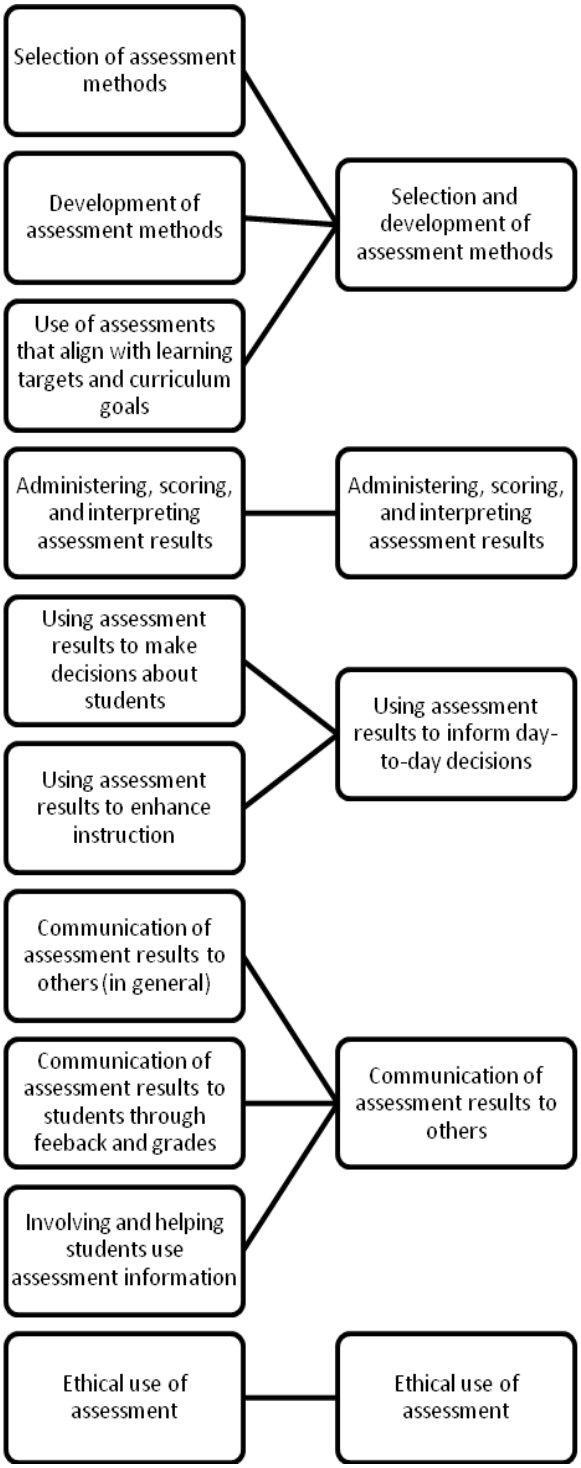
This research study does not require any additional interaction with the UMCIRB unless there are proposed changes to this study. Any change, prior to implementing that change, must be submitted to the UMCIRB for review and approval. The UMCIRB will determine if the change impacts the eligibility of the research for exempt status. If more substantive review is required, you will be notified within five business days.

The UMCIRB office will hold your exemption application for a period of five years from the date of this letter. If you wish to continue this protocol beyond this period, you will need to submit an Exemption Certification request at least 30 days before the end of the five year period.

The Chairperson (or designee) does not have a potential for conflict of interest on this study.

Appendix B

Diagram of hypothesized assessment areas for the STAP



Appendix C

Statement to expert judges

Purpose of the Study:
The following items will be used in a scale that will measure teacher assessment literacy. This scale is being created as part of a study for my school psychology graduate thesis. The purpose of the study is to develop a scale of teacher assessment literacy that measures teachers' skills in multiple areas of assessment literacy mentioned in the literature. Although this instrument is just being developed, I hope that it can ultimately aid school psychologists in consultation with teachers and guide school psychologists toward being better able to assist teachers with assessment issues. After reading the literature extensively, I believe that assessment literacy appears to be a 5-component construct. Although I may be wrong, this is the hypothesis I will be testing for this study.

Your Task as an Expert Judge:
There are 64 items (about twice as many as I expect to be in the final scale). For each item, you will be asked to indicate which of the five hypothesized components you believe the item belongs to, how confident you are that the item belongs to that component, how relevant you believe the item is to that component, and whether or not you believe the item is integral to the construct of teacher assessment literacy. The choice of components is a "forced choice" question. If you believe the item is related to more than one component, please make that note in the comment box. You may also include comments or suggestions for improvement as necessary. In addition, the four questions for each item must be completed before you move on to the next question. This is to ensure that I have all the information I need to apply decision rules regarding the deletion of items once I receive your feedback. Items will be deleted, added, or revised based on your feedback.

My definition of assessment literacy and a description of each component is displayed on the next page. A description of the components will be displayed with each item so that you may view them as you complete each item.

The survey should take approximately 25-35 minutes to complete. Once the survey is opened, you may save and continue at at later time.

Please complete this survey within four weeks (by January 10th, 2012).

Your feedback will be essential in assisting in the development of this scale. I want to thank you in advance for you time and feedback. It is valued and greatly appreciated!

Assessment literacy- "the possession of knowledge about the basic principles of sound assessment practice" (Paterno, 2001). For example, assessment literate educators will know what is being assessed, why it is being assessed, how best to assess the achievement of interest, what can go wrong, and how to prevent problems (Stiggins, 1995). In addition, teachers should be able to apply this knowledge to both formative and summative assessment.

Hypothesized Components:

1) Selection and development of assessment methods
Teachers should be able to select and develop appropriate methods and instruments for a variety of student needs. This includes the selection of multiple methods and strategies for each assessment. Teachers should also select methods that are aligned with standards and curriculum goals.

2) Administering, scoring, and interpreting results
Teachers should be able to administer both formative and summative assessments, be able to score the results, and be able to interpret these results.

3) Using assessment results for day-to-day decisions
Teachers should be able to use results of both formative and summative assessment to make decisions about both students and their instructional methods.

4) Communication of results to others
Teachers should be able to effectively communicate results to parents, students, and other educators. This includes communication through feedback and grades. Teachers should also be able to communicate results to students in a way that allows them to be involved in educational decisions.

5) Ethical use of assessment
Teachers should use assessment in an ethical manner and be able to recognize when it is not being used in an ethical manner.

Appendix D

Items by hypothesized component

Selection and development of assessment methods

3. Choosing an assessment method for a specific purpose, relating to an individual student
11. Selecting multiple methods of assessment (i.e. formal tests, in class observations, etc)
15. Creating assessments that accommodate the needs of all my students
17. Determining if an assessment is aligned with required standards (i.e. state or district curriculum goals)
18. Knowledge of which externally produced assessments are current and available
27. Developing assessments with different formats (i.e. multiple-choice, fill-in-blank, short answer)
29. Sampling from the domain defined by learning goals to write assessment items

Administering, scoring, and interpreting results

2. Seeking assistance when I am unsure of how to score an item
13. Interpreting summary scores reported with standardized test results (i.e. mean, percentile rank, standard scores).
14. Administering progress monitoring assessments
19. Administering standardized assessments (i.e. standardized achievement tests)
25. Interpreting criterion-referenced scores
26. Understanding why standardized administration is necessary to interpret results of standardized tests

Using results to inform day-to-day decisions

4. Using the results of formative assessment to adjust the content of my lessons
6. Using assessment results to appropriately group students for instruction
9. Using results of summative assessments to adjust future lesson plans
22. Using assessment information to develop an instructional plan for a student
23. Using progress monitoring results to adjust instruction
24. Using assessment results to identify students with similar needs

Communication of results to others

1. Explaining assessment results clearly to parents
7. Selecting appropriate methods for reporting results to parents, in addition to grades
8. Explaining results to other educators for the purpose of assisting with placement decisions
16. Explaining to parents how assessment results are used to make decisions about their children
21. Communicating the results of assessment to students in a way that they can understand
30. Explaining to students how results will be used to assign grades

Ethical use of assessment

5. Adhering to the bounds of confidentiality regarding assessment results
10. Knowing the consequences of unethical use of assessment
12. Recognizing inappropriate use of assessment
20. Recognizing when assessment results are being used inappropriately by others
28. Identifying my own legal responsibilities in regard to assessment

Appendix E

STAP cover sheet

The following packet contains the Scale of Teacher Assessment Practices (STAP), which is a scale on teacher assessment practices in the classroom that has been created and is being used for my master's thesis. The purpose of the study is to create and analyze a scale that measures multiple aspects of teachers' assessment practices in the hopes that the scale can help both teachers and school psychologists better understand how teachers use assessment. Your responses will aid in the analysis of this scale. Participation is voluntary, and by completing the STAP, you give permission for your responses to be used for research purposes.

Because the scale is only in initial development and accurate answers are essential to the analysis of the scale, it is very important that you answer all questions honestly. Your responses will remain confidential, and they will only be seen by myself and others directly involved in the study (e.g. my thesis advisor). When you have completed the scale, please hold on to it, and I will return to collect it.

In addition, please initial or check off the statements below to indicate that you have read and abided by each statement when completing the scale. Thank you so much for your time and cooperation, and please contact me if you have any questions or concerns!

_____
Catherine Cruess
School Psychology MA/CAS Candidate
East Carolina University
cruessc10@students.ecu.edu

_____ I have answered all questions honestly, and to the best of my ability.

_____ I have given myself an adequate amount of time to answer each question.

_____ I have given each question an adequate amount of attention, in order to read each question thoroughly before answering.

_____ I have answered all questions myself, without the assistance of others.

Appendix F

Scale of Teacher Assessment Practices (STAP)

Demographic Information

Directions:
This scale addresses issues in applying assessment practices in the classroom. Responses to these items will remain confidential. Participation is voluntary, and by completing the STAP, you give permission for your responses to be used for research purposes. Please fill in the following demographic information, or circle the appropriate answer when choices are provided.

1. Age: _____

2. Gender: M F

3. Number of years teaching: _____

4. Highest level of education: Bachelor's degree Master's degree Ph.D Ed.S

5. Grade level(s) currently taught: _____

6. Subject(s) currently taught: _____

7. Have you ever taken a course in measurement? Yes No (If yes, please continue)

   How many courses have you taken? 1-2 More than 2

Directions:
This scale addresses issues in applying assessment practices in the classroom. Responses to these items will remain confidential. Participation is voluntary, and by completing the STAP, you give permission for your responses to be used for research purposes.

There are 30 items relating to assessment practices that may be applied in the classroom. Each item is followed by a scale ranging from 1 (very low) to 5 (very high). Please estimate the **level of your skills** with each practice and circle the appropriate number. For example, if you feel that your skills are "Very High" with regard to "Explaining assessment results clearly to parents" then you would circle number 5. In contrast, if you feel that your skills relating to this assessment practice are "Very Low" then you would circle number 1. If you do not engage in a particular practice please circle "n/a".

	Very Low	Low	Acceptable	High	Very High	Not Applicable
1. Explaining assessment results clearly to parents	1	2	3	4	5	n/a
2. Seeking assistance when I am unsure how to score an item	1	2	3	4	5	n/a
3. Choosing an assessment method for a specific purpose, relating to an individual student	1	2	3	4	5	n/a
4. Using the results of formative assessment to adjust the content of my lessons	1	2	3	4	5	n/a
5. Adhering to the bounds of confidentiality regarding assessment results	1	2	3	4	5	n/a
6. Using assessment results to appropriately group students for instruction	1	2	3	4	5	n/a
7. Selecting appropriate methods for reporting results to others, in addition to grades	1	2	3	4	5	n/a
8. Explaining results to other educators for the purpose of assisting with placement decisions	1	2	3	4	5	n/a
9. Using results of summative assessments to adjust future lesson plans	1	2	3	4	5	n/a

	Very Low	Low	Acceptable	High	Very High	Not Applicable
10. Knowledge of the consequences of unethical use of assessment	1	2	3	4	5	n/a
11. Selecting multiple methods of assessment (e.g., tests, observations)	1	2	3	4	5	n/a
12. Recognizing inappropriate use of assessment	1	2	3	4	5	n/a
13. Interpreting summary scores reported with standardized test results (e.g., mean, percentile rank)	1	2	3	4	5	n/a
14. Administering progress monitoring assessments	1	2	3	4	5	n/a
15. Creating assessments that accommodate the needs of a variety of students	1	2	3	4	5	n/a
16. Explaining to parents how assessment results are used to make decisions about their children	1	2	3	4	5	n/a
17. Determining if an assessment is aligned with required standards (e.g., state or district curriculum goals)	1	2	3	4	5	n/a
18. Knowledge of which externally produced assessments are current and available	1	2	3	4	5	n/a
19. Administering standardized assessments (e.g., standardized achievement tests)	1	2	3	4	5	n/a
20. Recognizing when assessment results are being used inappropriately by others	1	2	3	4	5	n/a
21. Communicating the results of assessments to students in a way that they can understand	1	2	3	4	5	n/a

	Very Low	Low	Acceptable	High	Very High	Not Applicable
22. Using assessment information to develop an instructional plan for a student	1	2	3	4	5	n/a
23. Using progress monitoring results to adjust instruction	1	2	3	4	5	n/a
24. Using assessment results to identify students with similar needs	1	2	3	4	5	n/a
25. Interpreting criterion-referenced scores	1	2	3	4	5	n/a
26. Understanding why standardized administration is necessary to interpret results of standardized tests	1	2	3	4	5	n/a
27. Developing assessments with different formats (e.g., multiple-choice, fill-in-blank, short answer)	1	2	3	4	5	n/a
28. Identifying my own legal responsibilities in regard to assessment	1	2	3	4	5	n/a
29. Sampling from the domain defined by learning goals to write assessment items	1	2	3	4	5	n/a
30. Explaining to students how assessment results will be used to assign grades	1	2	3	4	5	n/a

Directions: The following 6 questions also address assessment practices. Please answer them in the following way: If you believe that you are highly skilled at applying the assessment practice, circle "5". If you believe that you are not skilled at applying the assessment practice, circle "1". If you feel that your response falls between a "1 and a "5", circle the appropriate number between "1 and "5".

	Not Skilled				Highly Skilled
1. Writing fill-in-the-blank/short answer questions	1	2	3	4	5
2. Using assessment results when developing lesson plans	1	2	3	4	5
3. Revising a test based on item analysis	1	2	3	4	5
4. Using assessments, such as classwork, to enhance my instructional delivery	1	2	3	4	5
5. Using assessment results to improve teaching and learning	1	2	3	4	5
6. Developing assessments based on clearly defined course objectives	1	2	3	4	5

*These questions were reproduced with permission of the copyright owners.