Utilizing a Theoretical Intervention to Examine Factors Influencing Teacher Efficacy toward Assessment and an Alternate Statistical Consideration for Program Evaluation

by

Shana Michele Shaw, B.A., M.Ed

Report

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

August, 2009

Utilizing a Theoretical Intervention to Examine Factors Influencing Teacher Efficacy toward Assessment and an Alternate Statistical Consideration for Program Evaluation

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: _____

(Gary D. Borich)

_____

(Marilla D. Svinicki)

Acknowledgements

I would like to thank my prospectus advisor and reader of this report, Dr. Marilla Svinicki, for her wise council at all stages of this project. I would also like to thank Dr. Gary Borich for his guidance through the program evaluation component of this report as well as my original prospectus committee members, Dr. Keenan Pituch, and Dr. Dan Robinson, for their feedback on an earlier draft of a portion of this report.

Utilizing a Theoretical Intervention to Examine Factors Influencing Teacher Efficacy toward Assessment and an Alternate Statistical Consideration for Program Evaluation

by

Shana Michele Shaw, M.A.

The University of Texas at Austin, 2009

SUPERVISOR: Gary D. Borich

In this research, a model of teachers' efficacy posed by Tschannen-Moran, Woolfolk-Hoy, and Hoy (1998) is considered with regard to teachers' use of standardized assessment data. This study is timely because teachers are expected to utilize standardized test scores, but they are often underprepared for this task. As a result of minimal experiences, teachers require in-service opportunities that develop their efficacy and knowledge toward standardized assessment. This proposal provides an opportunity for such experiences, and assesses the impact of a professional development activities designed to foster teachers' assessment efficacy and knowledge. Lastly, for considerations pertaining to program evaluation, this report will explore the relevance of using hierarchical linear modeling (HLM) as an alternative statistical procedure.

Table of Contents

Introduction

In 1998, Tschannen-Moran, Woolfolk-Hoy and Hoy traced the development of teacher efficacy back to its theoretical and empirical roots. By their account, the construct was born with the publication of the RAND report (1966, *cf* Tschannen-Moran et al., 1998) which described the findings of two items geared toward assessing teachers' perceptions of their influence on student achievement. The RAND items were theoretically derived from Rotter's Locus of Control Theory (Tschannen-Moran et al., 1998). From Rotter's perspective, feeling capable of controlling or affecting students' achievement formed the basis of teacher efficacy.

Another viewpoint on teacher efficacy resulted from Bandura's Social Cognitive Theory, particularly his concept of *self-efficacy* (Tschannen-Moran et al., 1998; Henson, 2002). According to Bandura (1977, 1993), self-efficacy beliefs are characterized by people's perceptions of their competence for a given task. When the construct was specifically applied to teachers' efficacy toward teaching, another conceptual thread was added to the development of teacher efficacy as a construct (Tschannen-Moran et al., 1998; Henson, 2002).

As a result of the simultaneous use of two theoretical bases, several researchers have lamented the empirical and theoretical problems inherent in teacher efficacy studies (Tschannen-Moran, Woolfolk-Hoy & Hoy, 1998; Henson, Kogan, & Vacha-Haase, 2001; Henson, 2002). These scholars have identified inconsistencies in this line of research which derive from measurement issues and theoretical confusion.

In spite of the perceived flaws in the construct (Tschannen-Moran, Woolfolk-Hoy & Hoy, 1998; Henson, 2002), the educational research community largely agrees that teaching efficacy predicts important teacher behaviors and student outcomes (Henson et al., 2001; Woolfolk-Hoy & Burke-Spero, 2005). For that reason, teacher efficacy is a viable construct to study in the context of an intervention geared toward improving teachers' efficacy toward assessment. To date, there have been few interventions geared toward influencing teachers' sense of efficacy and no interventions attempting to investigate the development of teachers' assessment efficacy with regard to large-scale assessment.

Therefore, the proposed research explores some of the current issues in teacher efficacy research by, as recommended by Bandura (1993), investigating the construct in the context of a specific teaching activity. This activity is appropriate given that teachers' utilization of standardized test data is a significant issue that currently affects educational practice. Through an intervention geared toward improving teachers' use of students' standardized test scores, the proposed research seeks to address teacher efficacy with regard to assessment.

Integrative Analysis

Positive Outcomes of Teacher Efficacy

Educational researchers identify teacher efficacy, or teachers' beliefs in their teaching ability, as a perception that is empirically related to teacher behaviors and student achievement (Woolfolk-Hoy & Burke-Spero, 2005). Woolfolk-Hoy and Burke-Spero (2005) posited that teachers' with high efficacy are more disposed toward experimenting with new instructional strategies and that they expend more energy in teaching. Empirical corroboration of these researchers' claims can be found in several studies. For instance, Stein and Wang (1988) demonstrated that teachers' implementation of an innovative mainstreaming program was related to their sense of efficacy for the implementation. These authors found that this efficacy was also related to the extent to which teachers incorporated the ideas posited by the program. On the other hand, research has demonstrated that teachers with low teaching efficacy were unlikely to integrate new, data-based strategies into their teaching practice (Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006). In another recent study, teachers' with low efficacy decreased their effort expenditure toward improving students' standardized test scores (Finnigan & Gross, 2007).

Selected student outcomes such as achievement and students' efficacy beliefs have also been linked reciprocally to teacher efficacy (Woolfolk-Hoy & Burke-Spero, 2005). Research investigating the means by which teacher efficacy affects student achievement has found that teachers with high efficacy attend more to the needs of lower-ability students (Ross & Bruce, 2007a). In addition, researchers posit that teachers with higher

efficacy often possess better classroom management skills and are better equipped to keep their students on-task (Tschannen-Moran et al., 1998). In terms of the relationship between teachers' efficacy beliefs and their use of student data, Foley (2007) posited that, when teachers feel capable of examining students' data, they make more informed pedagogical decisions which enhance student achievement. Since teachers with an elevated sense of efficacy seem to be better equipped to handle teaching challenges and tasks, considerations of how teacher efficacy develops, and whether it is amenable to change, are worthwhile.

Conceptions of Teacher Efficacy

Teacher efficacy, as previously mentioned, was derived from two theoretical bases: Rotter's *locus of control* theory and Bandura's concept of *self-efficacy* (Tschannen Moran et al., 1998). For a couple of reasons, Bandura's conception of self-efficacy provides a superior theoretical foundation for the current research. First, locus of control, as conceptualized by Rotter, is characterized by the perception of outcomes as internally or externally controlled (Tschannen Moran et al., 1998). This theory as related to teacher efficacy is reflected in the first item used in the RAND research which reads, "'When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment" (Henson, 2002). Bandura's self-efficacy is different from locus of control in that it is a person's future-oriented belief in his or her ability in a particular domain (Bandura, 1977, 1993). In the current context, Bandura's concept is more effective because a person may perceive that

he or she can control an outcome, but that person may not be secure in his or her ability to complete the requisite task (Ross et al., 1996; Tschannen-Moran et al., 1998). Put another way, Guskey and Passaro (1994) stated, "Individuals may believe that certain behaviors will produce particular outcomes, but if they do not believe they can perform the necessary actions, they will not initiate the relevant behaviors or, if they do, they will not persist in those behaviors" (p. 629). The current proposal seeks to determine what influences teachers' persistent beliefs in their ability to utilize standardized test data which falls more in line with Bandura's conception of self-efficacy.

Second, as reported by Tschannen-Moran et al. (1998), self-efficacy tends to be a robust predictor of behavior, while locus of control has weak predictive power. Consequently, in the current proposal, teacher efficacy will follow the advice of several researchers (Tschannen-Moran et al., 1998; Henson, 2002) and follow the line of research advocating for the use Bandura's self-efficacy construct as it relates to teacher efficacy rather than Rotter's locus of control variable. Since a goal of the proposed research is to facilitate improvements in teachers' efficacy, investigations of factors that may influence efficacy development are warranted.

Sources of Efficacy Information

The sources of efficacy beliefs posited by Bandura (1977) include *mastery experiences, physiological states* and *emotional arousal, vicarious experiences,* and *verbal persuasion*. *Mastery experience*, arguably the most influential source of information, is characterized by one's experience with a specific task. According to

Bandura (1977, p.191), "Cognitive events are induced and altered most readily by experience of mastery arising from effective performance". Frequent exposure to an activity results in the most salient changes in efficacy – when successes occur often, efficacy information becomes more positive (Labone, 2004). Conversely, frequent disappointments can negatively affect efficacy beliefs, especially when these beliefs are in their infancy (Labone, 2004).

Physiological states constitute another source of information from which people gather efficacy information. Elevated levels of physical arousal, whether the relative visceral experience is positive or negative, is a memorable experience in the development of efficacy beliefs. For example, Woolfolk-Hoy and Burke-Spero (2005) demonstrated that working too hard diminished teachers' efficacy, probably due to physical tiredness.

A third source of efficacy information, *vicarious experience*, is most informative when teachers observe their same-ability colleagues performing a teaching task (Labone, 2004; Ross & Bruce, 2007a). With regard to teachers' efficacy toward assessment, teacher efficacy has been developed vicariously when teachers observe the standardized test performance of comparable schools (Mandinach, Rivas, Light, Heinze, & Honey, 2006) and by watching low-performing districts make improvements by utilizing test data (Armstrong & Anthes, 2001).

The final source of efficacy information, *verbal persuasion*, is also dictated by involvement with others (Bandura, 1977). As indicated by Labone (2004), verbal persuasion from a credible informant can be especially powerful when teachers have pre-existing, damaged efficacy beliefs. Criticism from colleagues is an example of damaging

verbal persuasion (Woolfolk-Hoy & Burke-Spero, 2005). A lack of verbal reinforcement can be as powerful as the presence of it; in other words, teachers' efficacy can be damaged when they are not commended for their performance (Woolfolk-Hoy & Burke-Spero, 2005).

According to Henson (2002), the four sources of efficacy information have not been targeted for sufficient empirical examination, leaving several questions unanswered. Among them, Henson (2002) acknowledged that, though mastery experiences seem to have the strongest influence on the development of efficacy beliefs, this view does not provide insight into what types or features of mastery experiences provide the most useful information. Further, Henson questions whether certain sources may be more informative at different developmental stages in one's career. She posited that preservice and inservice teachers are affected differentially by the same information (Henson, 2002). In posing the question of how efficacy information may affect teachers differently according to their career stage, Henson tapped into an area of interest for many researchers who investigate the construct of teacher efficacy.

The Development of Teacher Efficacy

Studies investigating teacher efficacy often contend that teacher efficacy develops early in a teacher's career, primarily during student teaching and the first five years in practice (Ross, 1994; Woolfolk-Hoy & Burke-Spero, 2005; Ross & Bruce, 2007a). After those early years, teachers' efficacy beliefs appear resistant to change (Labone, 2004; Woolfolk-Hoy & Burke-Spero, 2005; Ross & Bruce, 2007a). Though research has

demonstrated that variation in teacher efficacy does occur as student teachers become practicing teachers (Woolfolk et al., 2005), the mechanisms underlying why or how teacher efficacy varies for teachers at different career stages have yet to be fully explored (Tschannen-Moran et al., 1998; Henson, 2002; Woolfolk-Hoy & Burke-Spero, 2005).

An interesting proposition for explaining some of the differences between preservice and inservice teachers' efficacy has involved a discussion of the salience of the task analysis among teachers (Tschannen-Moran et al., 1998; Henson, 2002). Several researchers have noted that teachers at the beginning of their careers rely more heavily on their analysis of the task for efficacy information while more seasoned instructors rely on their own experiences, especially if their prior experiences have occurred in similar contexts (Tschannen-Moran et al., 1998; Henson, 2002, Henson et al., 2002). Other researchers have noted the opposite effect; that is, experienced teachers may rely on the features of a task more than inexperienced teachers because these important task features become more noticeable and informative with practice (Ross et al., 1996). Either way, this research implies that task analysis is an important feature to consider in teacher efficacy research. Given this importance, survey instruments measuring teacher efficacy should incorporate items that are task-specific. Otherwise, the instrument might be too global to accurately assess changes in teacher efficacy.

A discussion of task analysis is especially pertinent if the task is as controversial as the one in the current research: utilizing standardized test data mandated by the *No Child Left Behind* (NCLB) Act. For instance, research has reported that teachers, under NCLB-related accountability practices, may feel as though they are being evaluated by their

students' standardized test performance which can lead to anxious or resentful feelings about standardized testing (Darling-Hammond & Wise, 1985; Haladyna, Haas, & Allison, 1998; Bernhardt, 2000). However, research on teachers' attitudes toward standardized testing implies that teachers may be becoming more amenable to utilizing the data that results from large-scale assessment.

According to Williams and Ryan (2000), in the early 1990s, teachers were more resistant to discussions of using test data. These authors contend that teachers' attitudes have shifted toward viewing standardized tests as information that could inform their teaching. Several empirical studies seem to reflect this shift. Educators have been concerned that standardized testing acts as a curriculum-narrowing mechanism, resulting in the neglect of significant instructional content (Darling-Hammond & Wise, 1985). Green and Stager (1986) demonstrated that teachers were wary of external testing, preferring their own classroom-based measures as indicators of student achievement. This wariness seemed, at times, to result from a lack of emphasis on measurement skills in teachers' preservice training and student teaching. Wise, Lukin, and Roos (1991) reported that, in preservice coursework and student teaching, future teachers were given the impression that it was not important to have skills in measurement and psychometrics.

Though not all recent studies report that teachers are supportive of using standardized test data (see Mulvenon, Stegman, & Ritter, 2005; Guskey, 2007), there is some evidence that an attitudinal shift may be occurring in favor of using test scores as information that informs instruction. A recent study determined that teachers agree that large-scale assessment can improve their teaching and their students' learning (Brown,

2002). Teachers, following instruction on how to utilize the available data for their teaching and given a supportive school-structure for data use, seem to be amenable to the idea that standardized test results can provide useable information (Chen, Salahuddin, Horsch, & Wagner, 2000; Protheroe, 2001; Ingram et al., 2004; Brunner et al., 2005; Wayman, 2005; Foley, 2007). In a recent, large-scale evaluation of teachers' data use in one school district, 92% of teachers agreed that utilizing data to inform pedagogical practices at their schools led to positive results (Wayman, Cho, & Johnston, 2007).

This pattern of improved attitudes toward standardized testing could be a result of the increasing use of data within schools to inform district- and school-level decision making. The practice, characterized by terms like *data-based decision making* or *data-driven decision making*, is typified by schools using multiple data points (e.g., standardized test scores, attendance records) to inform instructional practice. Researchers concerned with data-based decision making and policy implementation in school systems often investigate school-level factors that affect the success of the implementation of data-based methods in schools (Armstrong & Anthes, 2001; Ingram et al., 2004; Lachat & Smith, 2005; Louis, Febey, & Schroeder, 2005; Datnow, Park, & Wohlstetter, 2007, p.5; Mokhtari, Rosemary, & Edwards, 2007; Thomas, 2008). One major factor that has been found to predict school personnel's attitudes toward standardized testing, and the effectiveness of data-based decision making, is a school's' prior standardized test performance (Monasaas & Endelhard, 1994; Jones et al., 1999). As a result of students' test performance, schools are publicly-accessible ratings or rankings; for example, in Texas, schools are rated as *Exemplary*, *Recognized*, *Academically Acceptable*, or

Academically Unacceptable based largely on students' performance on the *Texas Assessment of Knowledge and Skills* (TAKS) exam. More information regarding these rankings is listed in Table 1.

Table 1.

Performance Standards for Texas' School-Rating Levels (2008)[1]

	Exemplary	Recognized	Academically Acceptable	Academically Unacceptable
TAKS Passing Rates	90% passing in all subject areas	75% passing in all subject areas	65% passing in Reading, Writing, Social Studies; 50% passing in Math; 45% passing in Science	Below the % rates specified by the "Academically Acceptable" rating levels

Given the impact a school's performance rating has been shown to have on the collective attitude toward standardized testing at each school, the currently proposed research takes school performance into account when investigating teachers' efficacy toward utilizing standardized test scores to inform their teaching practices. The construct of teacher efficacy, thus far, has been described as a unilateral construct; however, it is widely regarded and measured as a variable consisting of two-factors.

Teacher Efficacy Factors

In educational research, there are commonly used instruments that capture information about underlying cognitive, motivational, or affective constructs. In teacher

[1] Information retrieved August 9, 2008 from http://www.austin.isd.tenet.edu/newsmedia/releases/?more=1619&lang

efficacy studies, this seminal instrument is the *Teacher Efficacy Scale* (TES) by Gibson and Dembo (1984). Developed through an extensive construct validation study, the TES revealed a two-factor structure to teacher efficacy (Gibson & Dembo, 1984). These researchers initially labeled the factors *Personal Teaching Efficacy* (PTE) and *Teaching Efficacy*, but the second factor has evolved to be called *General Teaching Efficacy* (GTE) (Ross, 1994). According to Gibson and Dembo (1984), PTE represented the "belief that one has the skills and abilities to bring about student learning" (p. 573) while GTE measured the belief in "any teacher's ability to bring about change (that) is significantly limited by factors external to the teacher, such as the home environment, family background, and parental influences" (p. 574).

These two constructs were said to correspond to two of the expectancies identified in Bandura's Social Cognitive Theory (1989): *efficacy expectations* and *outcome expectations* (Gibson & Dembo, 1984). *Efficacy expectations* represent a person's belief that he or she can perform the actions necessary for a given task while *outcome expectations* represent the outcomes a person expects to arise after the performance of a task (Bandura, 1989; Tschannen-Moran et al., 1998; Henson, 2002). Tschannen-Moran et al. (1998) made the distinction that people's efficacy expectations partially inform their outcome expectations. Several authors have contended that, while measures of teaching efficacy should attend to teachers' outcome expectancies, a measure of teacher efficacy is not complete without an analysis of the task or context (Tschannen-Moran et al., 1998; Henson et al., 2002). No such task analysis component exits within the TES. Another criticism of the TES has been that the construct confusion in the instrument results from

the use of the RAND items, which were based on Rotter's Locus of Control Theory (Tschannen-Moran et al., 1998; Henson, 2002). A question arises, then, about the consistent two-factor structure revealed in the TES: what do these factors represent?

In a study geared toward uncovering the underlying constructs of the TES, Guskey and Passaro (1994) conducted a construct validity study on the measure. They observed that all of the items measuring the Personal Teaching Efficacy construct were positively worded and utilized the personal pronoun "I". Conversely, the General Teaching Efficacy items all used negative wording and the referent "teachers". By re-wording the items in the subscales, these authors determined that the resulting factors in the TES actually measured an internal vs. external orientation, which is in line with Tschannen-Moran et al.'s (1998) criticism that the TES derives from Rotter's Locus of Control Theory rather than Bandura's self-efficacy construct. Guskey and Passaro (1994) attributed the problems with the TES to possibilities including a mismatch between theory and operationalization, or that the measure was too global and not sufficiently specific to a domain (Guskey & Passaro, 1994).

As a counter to the operationalization of teacher efficacy posed by Gibson and Dembo (1984), Tschannen-Moran et al. (1998) posited a model suggesting that a valid measure of teacher efficacy includes both an assessment of personal teaching competence and an analysis of the task in terms of resources and constraints that are present in specific teaching contexts. Their model, consequently, argued that the two factors of teacher efficacy are *task analysis* and *personal teaching competence* rather than GTE and PTE.

According to Tschannen-Moran et al. (1998), *task analysis* bears some similarity to GTE in that it represents an analysis of the task based on whether teachers in general are equipped for the activity. It is distinct from GTE in that it includes elements specific to a particular teaching activity, and is not focused solely on barriers to effective teaching. An analysis of the teaching task, according to these authors, is informed by a teacher's assessment of factors that make teaching challenging pitted against the teacher's assessment of the available resources (Tschannen-Moran et al., 1998; Goddard, Hoy, & Woolfolk-Hoy, 2000). In order to analyze the requirements for a specific task, teachers should have an understanding of the complexity of task requirements (Tschannen-Moran et al., 1998). Gipps (1994) hypothesized that, when teachers are aware of the explicit requirements and goals of a task, their motivation is not compromised by a lack of understanding of the desired outcomes. As previously mentioned, analyses of tasks may differ for teachers at different career stages (Tschannen-Moran et al., 1998; Henson, 2002).

The second factor posited by the Tschannen-Moran et al. (1998) model relates to a teacher's personal feeling of teaching competence and is similar to Gibson's and Dembo's (1984) *personal teaching efficacy* (PTE) factor. *Personal teaching competence* is characterized by judgments of one's own teaching abilities leveraged against perceived personal weaknesses in a specific teaching activity. An important distinction between personal teaching competence and personal teaching efficacy relates to the former as a judgment of current abilities while the latter consists of perceptions of future outcomes (Henson et al., 2002). In terms of the sources of efficacy information, personal teaching

competence seems to be most informed by evidence of mastery experiences with a task (Tschannen-Moran et al., 1998; Henson et al., 2002).

Though previous research has lent some insight into the factors that influence teacher efficacy in a specific context, further research is needed to understand what role task analysis and personal teaching competence play in the formation of teacher efficacy (Tschannen-Moran et al., 1998; Tschannen-Moran & Woolfolk-Hoy, 2001; Henson et al., 2002). As yet, only a couple of studies have assessed the relative contributions of task analysis and personal teaching competence to teaching efficacy within a particular context. This could be due, in part, to the lack of a well-established teacher efficacy instrument that separately measures task analysis and personal teaching competence. The instrument that was developed in support of the Tschannen-Moran et al. (1998) model, the *Teachers' Sense of Efficacy Scale,* does not have items that distinguish task analysis from personal teaching competence. Instead, their instrument measures personal teaching competence and task analysis in tandem in three specific teaching contexts: *efficacy for instructional strategies, efficacy for classroom management*, and *efficacy for student engagement* (Tschannen-Moran & Woolfolk-Hoy, 2001). While this measure supports their contention that teacher efficacy is domain-specific, it does not allow for an exploration of the separate contributions made by the two-factors in their model (illustrated in Figure 1).

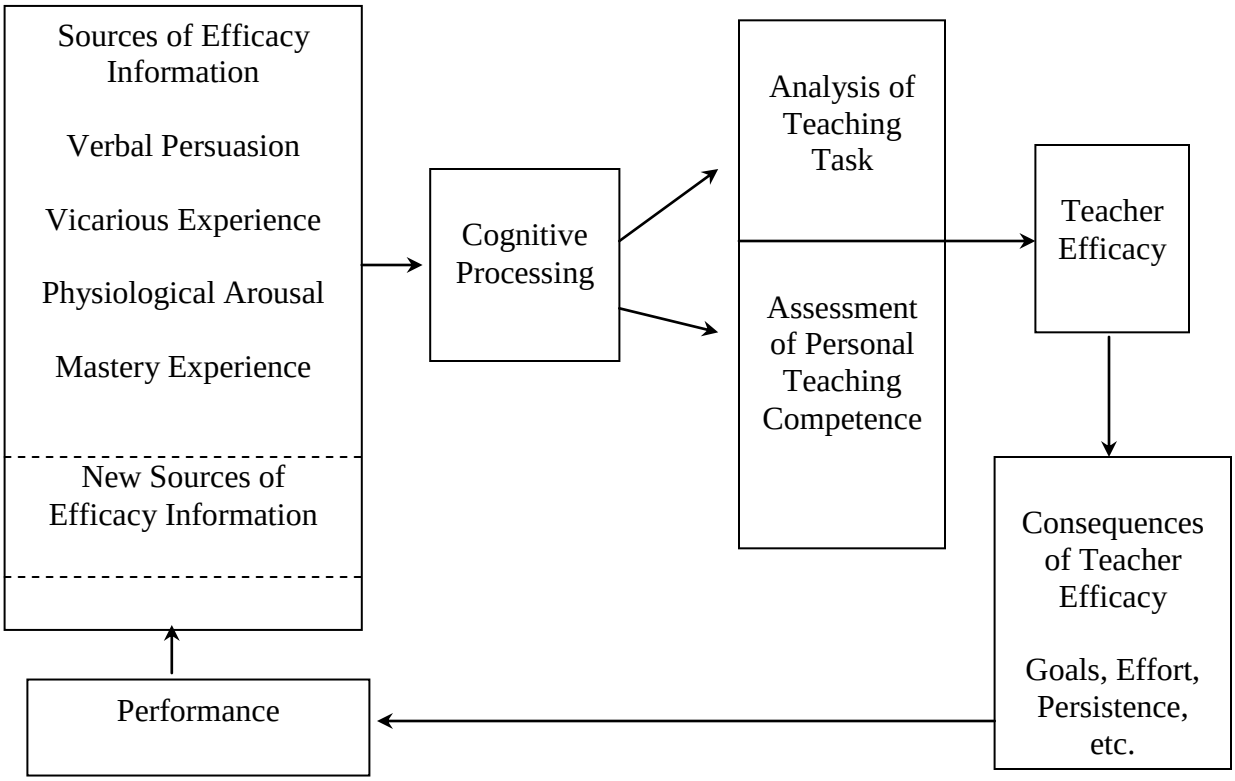


Figure 1.

Model of Teacher Efficacy posited by Tschannen-Moran et al. (1998, p. 228).

Though a teacher efficacy survey assessing task analysis and personal teaching competence has not been developed by Tschannen-Moran et al. (1998), other researchers have designed measures that evaluate these factors. Goddard et al. (2000) developed the *Collective Teacher Efficacy Scale* (CTES), and included both positively- and negatively-worded items designed to assess task analysis in teachers' assessments of collective teacher efficacy. At present, these items likely represent the most reliable measure of task analysis in a teaching efficacy instrument. The overall internal consistency measure

of reliability for the CTES was high ($\alpha = 0.96$). As such, these items will be adapted for use in the currently proposed research.

In another study examining the contribution of task analysis and personal teaching competence to teacher efficacy, Henson et al. (2002) developed methods to measure both of these factors as they relate to teacher efficacy. First, these authors developed the *Means-End Teaching Task Analysis* (METTA) to measure teachers' task analysis process. This measure consists of context-specific case studies and accompanying response sections that were designed to assess the extent to which task features facilitated or hindered teachers' ability to teach. These authors cited partial evidence supporting the score-related validity of the METTA. The reliability coefficients for the METTA's three task-specific subscales were 0.68, 0.70, and 0.67.

To measure the *personal teaching competence* component of teacher efficacy, Henson et al. (2002) adapted the personal teaching efficacy subscale of the TES. They felt confident in this adaptation because, like several other researchers, these authors found a three-factor solution to teacher efficacy as measured by the TES. Specifically, they found that the GTE subscale appeared to measure one factor while the PTE subscale actually measured two factors. The two-factor solution of the PTE subscale was posited to be a result of the manner in which items are worded. Items within the PTE subscale employ either a current-orientation or a future-orientation which means that some items assess how teachers currently feel about their abilities while others require teachers to postulate how they will perform in the future. This current- vs. future-orientation has important implications for the measurement of these concepts because it represents the

major distinction between Gibson's and Dembo's (1984) concept of personal teaching efficacy and the Tschannen-Moran et al. (1998) conception of personal teaching competence. Therefore, as a result of their findings, Henson et al. (2002) contended that the two factors in the PTE subscale measure both *personal teaching competence*, as conceptualized by Tschannen-Moran et al. (1998) and *personal teaching efficacy*, a more future-oriented factor of that is closely aligned with Bandura's conception of self-efficacy. As a result of this finding, they separated the original PTE subscale into two distinct subscales (measuring personal teaching competence and personal teaching efficacy). A low correlation between these new subscales supported the split (r=0.189). The coefficient alphas for the personal teaching competence and the personal teaching efficacy subscales were 0.70 and 0.60, respectively (Henson et al., 2002). Following the separation of these two factors, these researchers theorized that teachers' scores on the personal teaching competence subscale would predict their scores on the more future-oriented personal teaching efficacy subscale. Their findings supported this hypothesis – personal teaching competence was found to significantly predict personal teaching efficacy. However these authors reported different findings with regard to task analysis. Based on their data, they concluded that, while personal teaching competence was predictive of overall personal teaching efficacy (as measured by the future-oriented portion of the PTE subscale), only limited support could be found for task analysis as a predictor of efficacy outcomes. They did not discount the importance of task analysis, however, and they concluded that future research should explore the contribution of task analysis to teacher efficacy (Henson et al., 2002).

The preceding paragraphs have presented research assessing the relative impacts of task analysis and personal teaching competence on teacher efficacy. This proposed study seeks utilize portions of the methods used in these studies in several ways. First, this study will utilize modified items from the Collective Teacher Efficacy Scale (CTES) to measure teachers' analysis of factors that inhibit or facilitate the use of standardized test data to inform their teaching. Second, following the example set by Henson et al. (2002), this study will use items from the TES to measure teachers' context-specific judgments of their personal teaching competence. Third, these context-specific judgments of task analysis and personal teaching competence will be investigated for their unique contribution to the more future-oriented perception of personal teaching efficacy, as measured by the TES. Finally, this study will take into account teachers' knowledge acquisition with regard to assessment concepts, and how that acquisition might affect their efficacy. This is important because, though links between knowledge gains and self-efficacy are common in the literature, these connections are made less often between knowledge and teacher efficacy. Perhaps this is due to the fact that teacher efficacy is usually considered in terms of behavioral outcomes (e.g., teacher goal-setting, teacher persistence) rather than with content-knowledge gains. Another possibility is that prior research on teacher efficacy interventions were driven by content that is difficult to measure (e.g., classroom management skills, instructional strategies). Fortunately, the content of the currently proposed intervention is easily measurable, and can therefore be assessed in terms of its relation to teacher efficacy toward assessment. In total, this study

represents a contribution to the small, but growing, body of literature concerned with affecting changes in teacher efficacy through interventions

Intervention Research in Teacher Efficacy

Interventions geared toward developing or changing efficacy beliefs are not common in the literature (Henson, 2001). One exception to this is evidenced in the recent research conducted by Ross and Bruce (2007a; 2007b). These researchers designed a professional development system geared toward addressing teacher's efficacy in order to engender changes in the construct. Studied within the context of mathematics education, Ross and Bruce (2007b) designed a professional development workshop to address three types of teaching efficacy: *efficacy for engagement*, *efficacy for teaching strategies*, and *efficacy for student management*. These authors utilized an adapted version of the *Teachers' Sense of Efficacy* scale (Tschannen-Moran & Woolfolk-Hoy, 2001). The scale was adapted to reflect the specific content of the professional development exercise (i.e., mathematics education reform).

There were several strengths of this research including that these authors utilized a randomized trial involving almost all of the teachers in one school district. Further, while not all of their results were statistically significant, they did find that the treatment groups receiving information geared toward improving teaching efficacy showed higher posttest scores on the teaching efficacy constructs. These authors did not, however, investigate how task analysis and personal teaching competence may have differentially informed teachers' efficacy. Given their choice of measurement instrument, the *Teachers' Sense of*

Efficacy Scale, they were constrained to measure task analysis and personal teaching competence in tandem. Further, there was no discussion of how teachers' career stage may have influenced the effectiveness of the sources of efficacy information contained in the professional development activities. Both of these issues are slated to be investigated in the current research.

In their seminal article, Tschannen-Moran et al. (1998) described some issues that are important to consider during interventions geared toward improving teachers' efficacy. First, they mentioned that verbal persuasion during teachers' acquisition of new skills is important for the development of teacher efficacy. Second, they stressed that, in order for teachers to have mastery experiences, they must be given opportunities to rehearse their new skills. These suggestions regarding verbal persuasion and mastery experience are geared toward increasing personal teaching competence while Tschannen-Moran et al. (1998) suggested other means by which task analysis may be facilitated. They stated that teachers need a thorough understanding of the complexity of task requirements and assistance in learning how to manage the subset of skills required for a teaching task.

Professional development interventions often focus on specific teaching activities that have proven difficult for educators. The most effective professional development activities focus on enhancing both teachers' content and pedagogical knowledge (Guskey, 2003). In an example of such an intervention, Ross and Bruce (2007b) focused on the implementation of a standards-based mathematics curriculum. This new program had the potential to undermine teachers' efficacy since the instructional approach was

unfamiliar and required knowledge that the teachers had not yet acquired. In these cases, when teachers are introduced to new programs, they are often asked to implement novel teaching strategies with which they may not be familiar. In these situations, it is important to consider teachers' beliefs about their abilities to utilize the new methods and to assess how they may analyze features of the task at hand. In other words, it is important to consider teachers' efficacy during professional development.

In the current proposal, teachers will be asked to participate in professional development activities related to the use of large-scale assessment data. The use of these data has been mandated by state and national governments, and has proven useful in improving student learning when utilized in conjunction with other types of data (Wayman & Stringfield, 2006). Standardized testing, however, has been found to challenge teachers' sense of professional knowledge and credibility in their abilities to assess their own students (Graham, 2005). Ideally, an intervention geared toward improving teachers' efficacy toward using these types of data would improve both their knowledge and their feelings of capability in using students' test scores as information. Eventually, interventions of this type could prove to be useful in bridging the educational gap that exists between national- and state-policy mandates and teacher practice.

Guskey (2003) identified factors that lead to effective professional development. As already discussed, he stated that activities should improve teachers' existing content and pedagogical knowledge. Additionally, he mentioned that professional development must be efficient, purposefully planned, and well managed. This call for organized professional development activities is important because, especially with complex

concepts such as those related to standardized test scores and data use in general, the activities can quickly lose their scope and purpose. To guard against this potential lack of focus in the current proposal, a limited number of assessment-related concepts have been chosen for the proposed professional development.

Description of Assessment Concepts Addressed in the Intervention

According to Wise et al., (1991), teachers spend about 33% of their time engaged in various types of assessment tasks, yet 47% of preservice teachers thought their measurement training was "somewhat or very inadequate". This phenomenon could be related to a lack of measurement emphasis in preservice training. Approximately 70% of states do not require coursework or a demonstration of measurement knowledge for teacher certification (Rudner & Shafer, 2002). Consequently, preservice training in measurement and assessment often leaves teachers ill-prepared to interpret students' score reports from standardized tests (Marso & Pigge, 1988; Cromey, 2000; Mulvenon et al., 2005; Mandinach, Honey, Light, & Brunner, 2008).

In a study examining specific areas of deficiency in analyzing score reports, Impara et al. (1991) found that teachers had difficulty interpreting percentile band performance profiles, grade-equivalent scores, norm-group number correct and normal curve equivalent scores. These difficulties were present even in the presence of a guide meant to aid teachers in score report interpretation. Another study asked teachers about their comfort with interpreting score report data. These researchers found that 59% of teachers did not feel had sufficient training in analyzing score reports (Supovitz & Klein, 2003).

A related issue concerns preservice teachers' preparation in statistics. Creighton (2001, p. xii) suggested that statistics courses are not geared toward the needs of teachers. He stated that statistics courses focus on inferential statistics rather than concepts of actual use to teachers (e.g., descriptive statistics, data-based decision making, program evaluation) (Creighton, 2001, p. xiv). In a series of studies conducted at the University of Texas in Austin, Confrey and her colleagues investigated the development of teachers' statistical knowledge through a series of intervention studies. One of their first published studies revealed that teachers had difficulty understanding graphic representations of data as well as concepts related to statistical variation (Confrey & Makar, 2002). In a later study, Confrey and her colleagues found that, due to an instructional intervention, teachers' understanding of concepts related to distribution and variation improved. They deduced that it was good practice to involve teachers directly in the analysis of student data (Confrey, Makar, & Kazak, 2004). In yet another study, Kazak and Confrey (2004) used actual score reports during a measurement course for preservice teachers. Their instruction focused on topics similar to those they studied in past research such as distributional concepts, relationship concepts, and probability issues in between-group comparisons. On the whole, teachers' pre to posttest scores on these concepts improved, and the authors noted that teachers' motivation toward learning about score reports increased. It is important to note that their conclusions about teacher motivation are anecdotal – they did not actually assess teachers' motivation.

When developing measurement and statistics courses, it may be important to consider that preservice teachers find it difficult to appreciate that they will need

measurement training in their future teaching. They do not yet have real students with actual test scores and preservice teachers are not exposed to the accountability pressures experienced by practicing teachers. Perhaps inservice teachers make a better audience for training on measurement, but as yet, professional development on data use for teachers is limited (Protheroe, 2001; Mandinach et al. 2008). An attempt to offer this training, provided by Chen and colleagues (2000), seemed to register positively with the teachers in their study. Feelings toward the professional development were summarized by one teacher who stated, "This is the first time someone has made a sincere effort to explain the test scores to us and treated us as real professionals. No one has bothered to do this before" (p.379). Currently, and certainly in the future, data use will be more and not less complex since data warehousing and statistical software programs are becoming more sophisticated and geared toward large-scale usage (Wayman & Stringfield, 2006).

To provide a guideline for what teachers should know about assessment, the *Standards for Teacher Competence in Educational Assessment of Students* (1990) were developed during a collaborative project including the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association. The intentions of the Standards were that they be used by teacher educators to guide instruction in preservice teacher training and in professional development workshops as well as to give teachers a rubric for judging their own proficiency in educational measurement. Most of the Standards are geared toward teachers' use of classroom assessment, but Standard Three is explicitly tied toward understanding of

standardized testing concepts such as percentile ranks, percentile band scores, standard scores, and grade equivalent scores.

Since their development, the Standards have been used in applied research. For instance, Impara and Plake (1996) devised a professional development system meant to improve school administrators' knowledge of concepts related to standardized testing. Though the study was not directed toward teachers, it made two important contributions useful for the current proposal.

First, these authors pointed out that the *Standards for Teacher Competence in Educational Assessment of Students* (1990), while helpful to educators as a guide, do not prescribe specific "knowledge, skills, and abilities (KSAs)" (p.14) needed to utilize assessment data. Therefore, these authors developed a matrix of assessment-related KSAs and associated tasks with which these KSAs would be useful. The KSAs related to standardized assessment include KSA 1 (understanding jargon on standardized score reports), KSA 9 (interpreting standardized score reports to fellow educators), and KSA 11 (interpreting standardized score reports to parents).

The second contribution these authors made was to prescribe specific activities that could be used in professional development to assist educators in gaining understanding of these concepts. These will be discussed in greater length in the methods section since they serve as the basis from which some of the proposed professional development activities were developed.

Teachers Use of Standardized Test Results to Inform Their Teaching

One of the central concerns schools have in the implementation of the NCLB mandate that student data should inform instruction is simple: how? The authors of NCLB did not suggest methods for turning data into actionable information (Rudner & Shafer, 2002, p. 44; Wayman, 2005). Therefore, once educators have the knowledge about how to understand data, it is important to provide practical suggestions for data use in informing instruction.

First, teachers often rely on their "gut-feelings" in making pedagogical decisions (Ingram et al., 2004; Confrey & Makar, 2005; Foley, 2007; Moktari, Rosemary, & Edwards, 2007). Data provides a source of information that can supplement, though not replace, teachers' intuition. For example, standardized test score analysis provides teachers with tangible evidence about specific content areas with which their students have struggled (Chen et al., 2000).

Another way that teachers could utilize standardized test scores is to differentiate instruction. Differentiation is characterized by providing students at different developmental levels with specific assignments that address their needs (Protheroe, 2001; Brunner et al., 2005; Kerr et al., 2006). This activity is similar to providing students with Individualized Education Plans (IEPs), and though it may be time consuming, IEPs can have a powerful impact on student achievement (Supovitz & Klein, 2003; Codding, Skowron, & Pace, 2005).

In addition, it may be possible that teachers, in reflecting on students' scores, may identify areas that were lacking among all the students in their class. Teachers could give

extra consideration to those topics in the future and work to identify any deficiencies in their presentation of particular concepts (Kerr et al., 2006). Several researchers have suggested thinking about one's teaching practice can result in positive student-related outcomes (Brunner et al., 2005; Foley, 2007). Though this type of reflection regarding teachers' practice is plausible, it may be an unrealized potential in educational settings. Wayman et al. (2007) reported that the data use was often centered on individual students' specific needs, and not on how teachers could use data to enhance their overall pedagogical practice.

Another way teachers could utilize standardized test data is to engage in goal-setting with pupils and their parents regarding a students' future progress (Supovitz & Klein, 2003; Brunner et al., 2005; Mandinach et al., 2008; Wayman et al., 2007). Involving students in their own educational development can have a powerful impact on their cognitive and motivational development (Black & Wiliam, 1998). Further, it is a teacher's responsibility to explain students' scores to parents, and doing so in a positive way by setting future goals could help assuage parents' anxiety about their child's performance.

The topic of standards-based, standardized testing is a sensitive one. In terms of the usefulness testing can have for informing instruction, educational researchers and practitioners fall somewhere along a continuum ranging from "test results should never be used to inform instruction" to "test results are great, let's use them". Many researchers fall in the middle of that spectrum, and those individuals often advocate for a modified use of standardized test scores with special attention given to the fact that they

are not and should not be the sole indicator of a student's achievement or future ability (Kazak & Confrey, 2004).

From the perspective of a moderate use of standardized test results, the instructional intervention proposed in the current study will seek to inform teachers about measurement concepts that are pertinent to their practice. Additionally, the teachers will use this knowledge to engage in educational interventions for their students.

Focus on the Intervention

As teaching becomes increasingly professionalized, professional development must not only integrate content – it must also support motivational mechanisms through which desired educational outcomes are achieved (Henson, 2001). If teachers are to implement new ideas in their classrooms, teachers must feel confident about their ability to impact student learning with their new skills (Wolfe, Viger, Jarvinen, & Linksman, 2007). Professional development exercises, therefore, are not sufficient if they simply focus on communicating new knowledge or skills. These programs should also be receptive to teachers' need for reinforcement and efficacy (Fritz, Miller-Heyl, Kreutzer, & MacPhee, 1995). Unfortunately, prior teacher efficacy studies have largely utilized nonintervention-based designs. As stated by Ross (1994), intervention research is required if the educational community is to know what role teacher efficacy plays in teachers' implementation of new instructional practices. The current study represents an attempt to respond the need for research in this area; in particular, it seeks to investigate the malleability of teachers' efficacy toward assessment resulting from their engagement

in a professional development intervention geared toward improving their pragmatic skills and usage of students' results. The following research questions will be addressed in this intervention:

1. The first research question addresses teachers' posttest performance on assessment concepts, and how this performance might be related to teachers' assessment efficacy. The following sub-questions explore these outcomes:
    a. Are teachers' posttest scores on measurement concepts related to their post-intervention scores on the personal teaching efficacy scale?
    b. Will membership in one of the groups (Task Analysis, Personal Teaching Competence, Control) result in significantly higher posttest scores on measurement concepts?
2. The second research question is concerned with teachers' posttest performance on efficacy measures, how this performance might be influenced by teachers' experience level, and how these efficacy measures are related. The following sub-questions explore these relationships:
    a. What is the relationship among measures of teaching assessment efficacy (i.e., task analysis, personal teaching competence, personal teaching efficacy)?
    b. Will membership in one of the groups (Task Analysis, Personal Teaching Competence, Control) result in significantly higher post-intervention scores on any of the three teaching efficacy measures?

c. Will teachers' experience level significantly interact with treatment groups resulting in an aptitude-treatment interaction between the experience level and treatment groups?

Proposed Research Study

Statement of Purpose

The study proposed in this paper focuses on the effects of a professional development intervention aimed at improving teachers' assessment efficacy and knowledge of assessment concepts. In particular, the current proposal seeks to address three areas of teacher efficacy as it has been conceptualized by Tschannen-Moran et al. (1998). First, this research attends to a perceived need in teacher training regarding assessment concepts and teachers' efficacy toward their understanding and use of students' scores on standardized tests. Second, this study seeks to determine whether teachers' assessment efficacy is bolstered either through an intervention geared toward improving teachers' analysis of assessment tasks or toward improving their personal teaching competence. Last, this research proposes to investigate whether teachers' level of experience will affect the degree to which these activities successfully engender changes in their efficacy toward assessment.

Participants

The participants in this study will be elementary school teachers from a large, southwestern school district serving over 80,000 students on 110 campuses[2]. Of the 110 campuses, there are 81 elementary schools. The sample will be compiled through a stratified random sampling process using two strata: school performance ratings and teacher experience.

[2] Information retrieved January 24, 2008 from: http://www.austinisd.org/inside/

As mentioned, schools' standardized testing performance has been found to influence teachers' attitudes toward assessment (Monasaas & Endelhard, 1994; Jones et al., 1999). Therefore, the 81 elementary schools will first be stratified on their schools' performance rating (i.e., Exemplary, Recognized, Academically Acceptable, Academically Unacceptable) based on the 2008 TAKS test scores. According to the district's website, only six of the elementary schools were listed as Academically Unacceptable which sets the maximum school cell size for the other three categories. Out of the remaining 75 elementary schools, six schools within each performance category will be randomly selected, resulting in four groups from six schools for a total of 24 schools.

Because one of the goals of this research is to investigate the impact of teacher experience on teacher's assessment efficacy, teachers in these schools will also be stratified by their experience level. Principals from the 24 schools will be contacted to aid in the enlistment of teachers for the professional development activity. The principals will be asked to enlist eight teachers from their schools to participate based on the level of the teachers' experience: four novice teachers (i.e., less than five years of experience) and four experienced teachers (i.e., over ten years experience). Table 2 contains an illustration of the stratification procedure.

Table 2.

Illustration of Stratification of Participants (n = 192)

	School Performance level				
Teacher experience	Exemplary	Recognized	Acceptable	Unacceptable	*Total*
Novice	24	24	24	24	*96*
Experienced	24	24	24	24	*96*
Total	*48*	*48*	*48*	*48*	*192*

Number of Teachers in each group: Cell Sizes

	Within Factors			
Between Factor	Time 1 (Pretest)		Time 2 (Posttest)	
	Experience (≤ 5 years)	Experience (≥ 10 years)	Experience (≤ 5 years)	Experience (≥ 10 years)
Task Analysis Group (T1) (*n* = 64)	32	32	32	32
Personal Teaching Competence Group (T2) (*n* = 64)	32	32	32	32
Control Group (C) (*n* = 64)	32	32	32	32

After the sample has been compiled, participating teachers will be randomly assigned to one of three groups: Task Analysis group (treatment group one), Personal Teaching Competence group (treatment group two), or Control group. Given a total sample size of 192 teachers, there will be 64 teachers in each of these groups, with equal representation of novice and experienced teachers from schools with varying levels of performance.

As previously mentioned, the teachers selected for participation in the current proposal will be elementary school teachers. In particular, this study will recruit only teachers who teach third through fifth grade students because TAKS testing does not begin until third grade. Elementary teachers were chosen for two reasons. First, unlike middle and high school teachers who teach only certain subjects, elementary teachers are responsible for all of the content each student must master. The instruction provided in the professional development activities, therefore, will be appropriate for all teachers and will not require differentiation based on the teachers' content expertise (Math, Reading, etc.). Second, research has shown that grade-level taught is not likely to impact whether teachers are capable of learning about assessment topics. For instance, in his research on fourth, eighth, and eleventh grade teachers' ability to understand score reports, Impara et al. (1991) demonstrated that grade level taught did not significantly impact whether teachers were able to interpret score reports.

Approval for this study will be sought through the Institutional Review Board (IRB) committees at both the university- and school district-levels. IRB approval will be sought for one full academic year. Informed consent regarding the purpose of the study will be obtained from all participants.

Materials

Teacher Efficacy Measures

In the current study, adaptations of the *Collective Teacher Efficacy Scale* (CTES) (Goddard et al., 2000) and the *Teacher Efficacy Scale* (TES) (Gibson & Dembo, 1984)

were made to estimate teachers' task analysis, personal teaching competence toward assessment, and personal teaching efficacy toward assessment. Alterations of teacher efficacy instruments are common in this line of research due to the specificity with which teacher efficacy should be measured. Notable adaptations of the TES include the *Science Teaching Efficacy Belief Instrument* (STEBI) and the *Dutch Teacher Self-Efficacy Scales* (*cf* Tschannen-Moran et al., 1998). These modifications set the precedent for allowing revisions to measures in order to suit the content in which teacher efficacy is being investigated.

Task Analysis Subscale. In order to measure task analysis, items from the CTES have been modified. The original, unmodified CTES appears in Appendix A. The following six items, measured on a six-point Likert-type scale (strongly disagree to strongly agree), represent the items that will measure task analysis (these correspond to items 11-16 on the CTES):

1. Students come to school ready for instruction based on their standardized test results.
2. Homelife provides so many advantages that students are bound to be able to excel in environments that utilize their standardized test results.
3. The lack of instructional materials and supplies makes utilizing students' standardized test scores to improve instruction very difficult (reverse coded).
4. Students here just aren't motivated to work on issues related to their standardized test performance (reverse coded).
5. The quality of school facilities here really facilitates the use of standardized test scores to improve instruction.

6. The opportunities in this community help ensure that I will be able to utilize students' standardized test scores to improve my instruction.

Personal Teaching Competence Subscale. Personal teaching competence will be measured by adapting items from the personal teaching efficacy subscale of the TES. This adaptation is similar to the one demonstrated in research by Henson et al. (2002) except that the modifications refer to teachers' personal teaching competence with regard to assessment. The unmodified PTE subscale used in Henson et al. (2002) is presented in Appendix B. These items, measured on a six-point Likert-type scale (strongly disagree to strongly agree), represent the items that will measure personal teaching competence (these correspond to items 1, 6, 9, and 10 on the TES):

1. When a student does better than usual on a standardized test, many times it is because I exerted a little effort.
2. When a student does better than usual on a standardized test, it will usually be because I found a better way of teaching the student.
3. When the standardized test scores of my students improve, it is usually because I found more effective teaching approaches.
4. If a student masters a section of a standardized test, this might be because I will have known the necessary steps in teaching the content of that section.

Personal Teaching Efficacy Measure. Personal teaching efficacy will also be measured by adapting items from the personal teaching efficacy subscale of the TES. As with research conducted by Henson et al. (2002), these items will measure teachers' future-oriented perceptions of their efficacy toward assessment. These items, measured

on a six-point Likert-type scale (strongly disagree to strongly agree), represent the items that will measure personal teaching efficacy (these correspond to items 7, 13, 15, and 18 on the TES):

1. When I really try to improve my students standardized test scores, I will be able to even with the most difficult students.
2. If one of my students becomes disruptive during a standardized test administration, I feel assured that I will know some techniques to redirect him/her quickly.
3. If one of my students did not pass a section of a standardized test, I would be able to accurately assess whether the test items were the correct level of difficulty.
4. If I try really hard, I will be able to improve the standardized test scores of even the most difficult or unmotivated students.

Assessment Concepts Measures

In order to measure teachers' conceptual skills in assessment, pre and post measures were developed from several sources. These references include the *Texas Assessment of Knowledge and Skills* (TAKS) Usage Manual (2008) and several educational measurement texts (Rudner & Schafer, 2002; Klein, 2005; Stiggins, 2007). Generally, the pre and post measures contain the same questions, but items with numerical answers were changed slightly to protect against answer recognition. Additionally, several of the items require teachers to identify concepts on recreated TAKS score reports. For instance, item 10 prompts teachers to refer to a TAKS score report to identify the content area in which a hypothetical student needs the most improvement. These pre and post measures of teachers' assessment skills are located in

Appendices C and D. An example of the accompanying TAKS score report used in the measurement of assessment knowledge is located in Appendix E.

Procedure

In this research study, two professional development interventions (i.e., Tasks Analysis and Personal Teaching Competence treatment groups) will be implemented by the researcher. In total, the three-week long professional development exercises are slated begin when TAKS scores are released to schools in the summer and to extend slightly into the beginning of the school year. Texas schools receive their students' TAKS scores in the first week of August which makes activities geared toward understanding and using these scores especially relevant at that time. Further, teachers have more time to devote to professional development in the summer than they do during the regular school year. It is important to note that this study is potentially taking a risk in asking teachers to continue their participation during the beginning of the school year. However, this might be necessary because teachers' efficacy toward assessment will likely remain unchanged unless they are given the opportunity to use their newly acquired skills to inform their instructional practice (Henson, 2001).

The first week will consist of pre-testing all participants on assessment concepts and the three aspects of teacher efficacy toward assessment, and for the treatment groups, instruction on assessment concepts. The second week will consist of review of the previous instruction as well as planning regarding the use of score results to design a beginning-of-the-year, teacher-based assessment on TAKS-related topics. The rationale

behind engaging the teachers in constructing their own assessments based on the TAKS results is three-fold. First, students take the TAKS test in March[3], so these scores may not reflect students' current understanding of pertinent content. Taken alone, these tests do not provide teachers with feedback that is timely enough to inform pedagogical practices in the classroom (Supovitz & Klein, 2003). By conducting their own beginning-of-the-year assessment of students' current functioning, these teachers are obtaining a more recent snapshot of students' readiness for instruction. Second, research has reported that teachers trust their own, teacher-developed assessments more than they trust large-scale assessment results (Mulvenon et al., 2005; Guskey, 2007). This activity lends them the opportunity to develop their own classroom assessments. Last, this activity will require teachers to engage in an examination of the broad content areas assessed by the TAKS test and to think of ways to connect those topics to their own classroom activities. According to Supovitz and Klein (2003), in creating their own assessments based on standardized test scores, teachers are reflecting on the connections between content standards, classroom instruction, and ultimately, student achievement. This practice of using data to reflect on instructional practice and content has been recommended by prior research (Brunner et al., 2005; Foley, 2007; Wayman et al., 2007) and is mandated by the NCLB Act (Rudner & Shafer, 2002, p. 44; Wayman, 2005). The final week in the professional development will be different for teachers depending on

[3] Information retrieved on August 9, 2008 from http://www.tea.state.tx.us/student.assessment/admin/calendar/2007_2008_revised_01_17_08.pdf).

their group assignments, but posttests on measurement concepts and teachers' efficacy will be administered to all groups at this time.

An important feature of the intervention is that the two treatment groups will both receive instruction on the same assessment concepts, but the delivery of the training will be different depending on whether they are in the Task Analysis group or the Personal Teaching Competence group. In this way, it may be possible to observe the separate influences of task analysis and personal teaching competence on a teachers' efficacy for assessment. In addition to the treatment groups, a waiting-list control group receiving no instruction on these concepts will be used for comparison purposes. Further explanation of the activities of each of the treatment groups and the control group follow in the paragraphs below.

Treatment Condition One: Task Analysis

As posited by Tschannen-Moran et al., (1998), teachers' knowledge of task features affects their efficacy toward a task. In the context of the current intervention, task analysis will be facilitated largely through verbal explications of the skills required to utilize assessment data. This will be done primarily through lectures on assessment concepts, followed by researcher-led demonstrations of how teachers could construct their own assessment items that cover the areas addressed on the TAKS test. This portion of the instruction will cover topics such as utilizing tables of specification for test construction, item writing and item analyses for classroom-based assessments. Teachers will be required to construct their own beginning-of-the-year assessments based on the results of the students who will be in their classes.

Additional instruction developed solely for this condition will explicitly focus on the potential environmental resources and barriers to using standardized test scores to inform instruction. Resources include products geared toward assisting teachers in modifying their instruction based on TAKS results. These products are released by entities such as the commercial testing company that publishes the TAKS and the Texas Education Agency (TEA). Barriers to implementation might include the lack of access to these necessary resources, the timeline of the release of TAKS results, and a lack of time needed to prepare the suggested materials. These resources and barriers will be discussed with participants in this condition. Overall, this condition will focus on the features of the task of using standardized test data, but will not engage in the teachers in having incremental mastery experiences in using these data as will be the case in the Personal Teaching Competence group.

Treatment Condition Two: Personal Teaching Competence

The Personal Teaching Competence treatment group will receive instruction geared toward increasing their personal teaching competence through opportunities to have mastery experiences in learning and applying the score results. Therefore, the sessions for the participants in this group will include activities that provide the teachers with repeated feedback on their mastery of these concepts. Following the initial instruction on assessment concepts, teachers will engage in several activities similar to the ones recommended by Impara and Plake (1996). For instance, the first activity will consist of having teachers summarize the test performance of a hypothetical student. Another activity will require teachers to identify topics for discussion during a parent-

teacher conference for the student. These teachers will then collaborate in groups to prepare a report on their summary and parent-teacher discussion which will be assessed by the author of this report. Feedback will be given to participants during a subsequent session.

Additionally, as will be the case with the Task Analysis group, teachers in this condition will be required to construct their own beginning-of-the-year assessments based on the results of the students who will be in their classes. The last session for this group only will be used to analyze the teacher-based assessment results and to further plan pedagogical activities for the coming year based on the findings of the teacher-based assessments. One suggestion for an activity includes having teachers create graphs on which they will track their students' progress toward learning standards throughout the school year (Supovitz & Klein, 2003).

Control Condition

The Control group will receive no immediate instruction regarding assessment concepts and their applicability for instruction; therefore, this group will be a waiting-list control group. This condition is essential because the treatment groups' results will be gauged against the control group's results to assess the effectiveness of the treatment. However, these teachers will have volunteered for a professional development on utilizing assessment for instruction, and therefore, will be compensated with instruction. Consequently, a separate, post-intervention workshop will be provided to the control teachers following the completion of data collection for the current research. Table 3 outlines the scheduled activities in the professional development.

Table 3.

Scheduled Activities for Treatment Groups and the Control Group

Group	Week 1	Week 2	Week 3	Post-Intervention
Task Analysis	Pre-testing; Assessment instruction	Review Assessment instruction; Test construction	Post-testing	*
Personal Teaching Competence	Pre-testing; Assessment instruction; Mastery activities	Review Assessment instruction; Test construction; Mastery Activities (e.g. summarize score reports)	Post-testing; View results of teacher-made tests; Instructional planning (e.g., make progress charts)	*
Control	Pre-testing	*	Post-testing	Assessment instruction

* indicate that no activity will take place for that group

Research Questions, Hypotheses, and Analyses

Research Question One: Knowledge Measures

The first research question addresses teachers' posttest performance on assessment concepts, and how this performance may be related to teachers' assessment efficacy. The following sub-questions explore these outcomes:

a. Are teachers' posttest scores on measurement concepts related to their post-intervention scores on the personal teaching efficacy scale?

b. Will membership in one of the groups (Task Analysis, Personal Teaching Competence, Control) result in significantly higher posttest scores on measurement concepts?

Hypotheses for Research Question 1

The following hypotheses are related to the first research question:

a. In the treatment groups, teachers' posttest scores on assessment concepts will be significantly related to their post-intervention scores on teacher efficacy as measured by the personal teaching efficacy subscale. This relationship is hypothesized to be a positive one; in other words, as teachers' knowledge of assessment concepts increases, so does their efficacy with regard to assessment. It is important to note, however, that the directionality of this relationship is not investigated in the current study; in other words, the researcher does not posit whether knowledge gains precede efficacy gains or vice versa.
b. Teachers in the treatment groups receiving instruction on measurement concepts will score significantly higher than teachers in the control group on the assessment concepts posttest. Also, teachers in the Personal Teaching Competence treatment group will significantly outperform teachers in the Task Analysis group on the posttest because they will be given opportunities to influence their mastery of their measurement skills through activities.

Analyses for Research Question 1

The first sub-question of the first research question asks whether posttest scores on the measurement concepts are related to post-intervention scores on the personal

teaching efficacy toward assessment subscale. To address this question, a Pearson product moment correlation (*r*) will be computed between these two variables to assess the magnitude and direction of the relationship between these variables. As mentioned above, the expected result is a significant, positive correlation coefficient.

In attending to the second sub-question addressed by the first research question, this study will utilize a 3 x 2 repeated-measures ANOVA with one between-factor with three levels (group assignment) and one within-factor with two levels (measurement of subjects' assessment knowledge). The dependent variable in this question is teachers' posttest performance on the assessment measure. For this question, only the between-factor (group assignment) will be analyzed for mean differences in posttest knowledge. Schematically, the repeated-measures design is represented in Table 4.

Table 4.

Repeated Measures Design (DV: Posttest Scores on Assessment Concepts)

	Within Factor	
Between Factor	Time 1 (Pretest)	Time 2 (Posttest)
Task Analysis Group (T1) (n = 64)		
Personal Teaching Competence Group (T2) (n = 64)		
Control Group (C) (n = 64)		

* Alpha will be set, a priori, at 0.01.

A repeated-measures design was considered appropriate for the current research given that teachers will be randomly assigned to conditions[4]. Additionally, a repeated-measures design is more robust in minimizing within-subject error variance than a standard ANOVA which increases the power of this study in detecting differences between treatment groups if, in fact, differences are present (Stevens, 1999). In terms of the assumptions of the repeated-measures design, it is possible that the independence of observations assumption will be violated. A violation of this assumption may occur in one of two ways: school membership or treatment group membership. First, though the sampling process employed a stratified procedure to protect against the confound of schools' test performance, other school-level factors such as leadership attitudes toward data use or school SES may differentially impact teachers from different schools. Additionally, teachers in the Personal Teaching Competence treatment group will be collaborating on several activities which may also result in a violation of this assumption. Since violations of the independence assumption can result in inflated Type I error rates (Stevens, 1999), the alpha level for this study has been set a priori at 0.01 as opposed to the less stringent 0.05. With regard to the assumption of multivariate normality, scores on the posttest will be assessed for skewness and kurtosis. Unless deviations are severe, which is not hypothesized in the current research, repeated-measures ANOVAs are robust to violations of multivariate normality (Stevens, 1999). Finally, with regard to the sphericity assumption, this assumption will also be explored, though there is no reason to

[4] Without random assignment, an ANCOVA design would have been more appropriate.

believe that there will be unequal variances between groups from the pretest to the posttest.

SPSS software will be used to analyze these data. The main effect for the between-factor will be assessed to determine whether there are significant differences in posttest scores due to group assignment. If an overall difference is found, a Tukey post hoc test will be used to conduct pairwise comparisons among the groups to determine where the difference is situated.

Research Question Two: Efficacy Measures

The second research question is concerned with teachers' posttest performance on efficacy measures, how this performance might be influenced by teachers' experience level, and how these efficacy measures are related. The following questions explore these relationships:

a. What is the relationship among the dependent measures (i.e., task analysis, personal teaching competence, personal teaching efficacy)?
b. Will membership in one of the groups (Task Analysis, Personal Teaching Competence, Control) result in significantly higher post-intervention scores on any of the three teaching efficacy measures?
c. Will teachers' experience level significantly interact with treatment groups resulting in an Aptitude-Treatment Interaction (ATI) between the experience level and treatment groups?

Hypotheses for Research Question 2

The following hypotheses are related to the second research question:

a. Teachers' posttest scores on these three teacher efficacy outcome measures will be related. The strength of the relationship between the personal teaching competence and personal teaching efficacy measures will be stronger than the relationship between task analysis scores and the other two efficacy measures. All relationships will be positive.

b. Teachers in the treatment groups will score significantly higher than teachers in the control group on the post-intervention administration on all of the teacher efficacy subscales. Teachers in the Personal Teaching Competence treatment group will have greater gains on the posttests of personal teaching competence and personal teacher efficacy toward assessment. This hypothesis is supported by Henson et al. (2002) who did not find that task analysis was predictive of personal teacher efficacy. Teachers in the Task Analysis group will have greater gains on the task analysis posttest than participants in the other groups.

c. An ATI suggests that certain treatments are better for certain participants based on their personal characteristics (Keith, 2006). In the proposed research, it is theorized that, as stated by Tschannen-Moran et al. (1998), less experienced teachers will rely more heavily on features of the task to inform their efficacy while more experienced teachers will rely more on their own personal experiences. Therefore, it is hypothesized that teachers' experience level will significantly interact with the treatment groups to produce differential outcomes on the post-intervention task analysis and personal teaching competence measures. Specifically, newer teachers in the Task Analysis group will have higher task analysis gains than more experienced

teachers. Further, more experienced teachers' efficacy in the Personal Teaching Competence group will improve more as measured by the personal teaching competence posttest.

Analyses for Research Question 2

The first sub-question of this research question is concerned with whether these three subscales measure related constructs. In order to determine the strength and direction of this relationship, Pearson correlations will be calculated between the three scales.

The second and third portions of this question will be investigated utilizing a repeated-measures MANOVA. This repeated-measures MANOVA will have three dependent variables: task analysis, personal teaching competence and personal teaching efficacy. This analysis will have one between-factor (group assignment) and two within-factors (posttest scores on teacher efficacy posttests and teacher experience). The design is represented in Tables 5.

Table 5.

Repeated Measures Design (DVs: Posttest Scores on Task Analysis, Personal Teaching Competence, Personal Teaching Efficacy)

	Within Factors			
Between Factor	Time 1 (Pretest)		Time 2 (Posttest)	
	Experience (≤ 5 years)	Experience (≥ 10 years)	Experience (≤ 5 years)	Experience (≥ 10 years)
Task Analysis Group (T1) (*n* = 64)				
Personal Teaching Competence Group (T2) (*n* = 64)				
Control Group (C) (*n* = 64)				

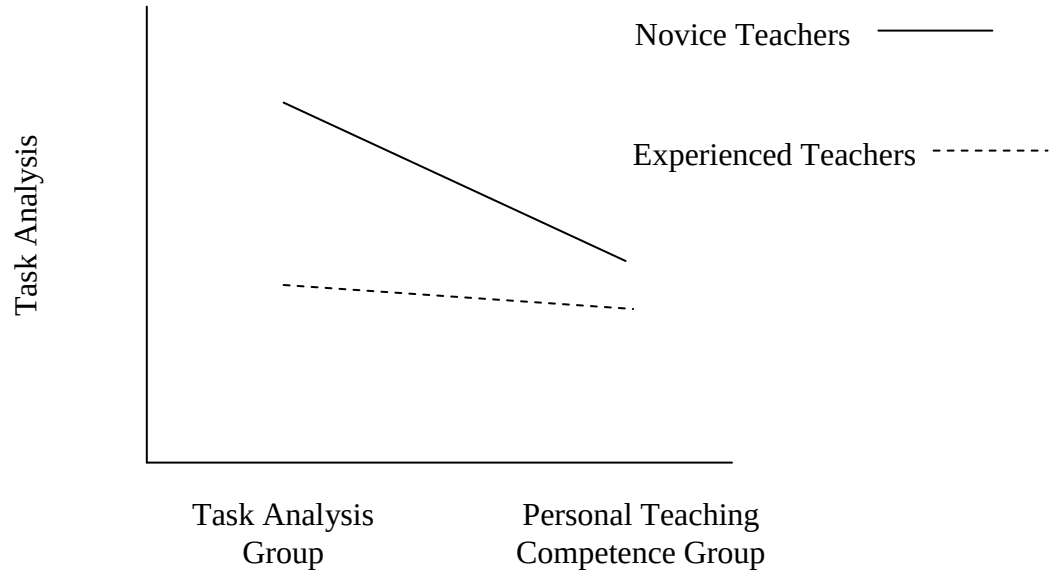
* Alpha will be set, a priori, at 0.01.

A repeated-measures MANOVA was appropriate in this case since an argument can be made that the three outcome measures are theoretically connected as demonstrated by model posed by Tschannen-Moran et al. (1998). If the overall test statistic (Wilk's Lambda) for the MANOVA is significant, separate univariate tests will be conducted to determine where the differences lie. Given the fact that three separate univariate tests will be used, and if differences are found, subsequent post hoc tests, the a priori alpha level was set at 0.01 to control for an inflated Type I error rate. Further, ATIs will be considered in the univariate tests with task analysis and personal teaching competence scores as dependent variables. This interaction will be considered between teacher experience (within factor) and group assignment (between factor).

Figure 2 displays an example of the hypothesized interaction between the teacher experience, group assignment, and task analysis as the outcome variable.

Figure 2.

Hypothesized ATI between treatment groups and teacher experience (DV: Task Analysis)



Additional Analyses

Reliability coefficients (internal consistency estimates) will be computed for each of the subscales used to measure teacher efficacy in the current research. This is appropriate for a couple of reasons. First, the measures used in this research to assess the components of teacher efficacy (task analysis, personal teaching competence for assessment, personal teaching efficacy for assessment) have been modified for use in the current study. Therefore, existent estimates of reliability associated with scores on these subscales would not apply to this research. Second, since reliability estimates are

sample-dependent indications of the degree of measurement error, educational researchers have recommended that these coefficients be reported, as common practice, in all studies utilizing survey methodology (Vacha-Haase, Kogan, & Thompson, 2000).

Discussion

Summary

The previous sections of this paper have identified the reasons why a study investigating teachers' efficacy toward the use of standardized assessment data to inform their instruction is viable, current, and necessary. The sheer lack of empirical research devoted to the topic lends credence to the need for further study (Zhang & Burry-Stock, 2003; Ingram, Louis, & Schroeder, 2004; Lachat, 2005). Further, as suggested by Impara, Plake, and Fager (1993), it is important to also consider teachers' beliefs and attitudes toward assessment, an area that has been neglected in empirical studies. Last, interventions geared toward teaching efficacy are rare in the research, though the ones that have taken place have shown that teaching efficacy is amenable to intervention (Ross & Bruce, 2007a; Ross & Bruce, 2007b). There is certainly no disputing the need to involve teachers in an intervention of this kind given their impact on practices within schools and student achievement.

Limitations

The limitations inherent in most social science research are present in the currently proposed study as well as some limitations particular to the context of this particular study. The general limitations include threats to internal and external validity, selection bias among participants, and the possibility of Type II errors due to strict alpha levels. First, in terms of threats to internal validity, though attempts were made to protect against potential confounds (e.g., using a sample stratified on school performance), it is

possible that extraneous, unmeasured variables could affect the outcomes of this study. Of particular concern to the internal validity in this study is whether the treatment groups have been appropriately designed to address teachers' task analysis or personal teaching competence. It is possible, for example, that teachers in these groups could respond to the delivery style as more lecture-based (Task Analysis group) or more collaborative (Personal Teaching Competence group). Unfortunately, no previous research has set a precedent for how to address these factors through an intervention. This study represents the first intervention-based test of the components of the Tschannen-Moran et al. (1998) model. Therefore, regardless of the findings, replications and modifications would certainly need to be made in future research studies. In terms of external validity, it is possible that the results of this study cannot be generalized beyond the current sample, and it may be likely that the results would not generalize to teachers in different parts of the country with different state-mandated regulations regarding standardized testing.

Another limitation concerns the potential for self-selection bias among participants. Given the nature of conducting research in practical pedagogical settings, the teachers will likely have to be volunteers who agree to participate in the professional development activities. Because it is not realistic to expect that teachers could be required to participate, this characteristic of the design was regarded as necessary. Ideally, the teachers who will self-select for participation will prove to be emissaries of the information to other teachers at their schools who may not have been eager or able to take part.

Another limitation is the possibility that Type II errors could occur as a result of the strict alpha levels used in the current study. This possibility was given much thought, because it would be unfortunate to expend such effort toward improving teachers' assessment efficacy and skills, and to find that the results were not significant due to a strict a priori alpha level. However, given that teachers' time and effort is so valuable, it is important that the effects of professional development activities such as these be supported by strong resultant data. For that reason, it was determined that a stringent alpha level that protected against Type I errors would be appropriate.

The limitations due to the context are related to the larger picture in which these activities take place. Teachers will return to schools that have varying attitudes and degrees of amenableness to data use. One of the most critical components supporting teachers' use of these types of data is that they be given training and time to improve their requisite skills and instructional strategies (Wayman et al., 2007). The current research represents an attempt to address the former need in terms of teachers' training, but it cannot provide teachers with the time and support that they will need to enact changes in the classroom. That support must come from their schools' and the district, and will only happen if these entities also share a vision for using standardized test scores and other types of data to support students' learning.

Addendum: Methodological and Statistical Considerations in Program Evaluation

The purpose of this addendum is to consider an alternate statistical procedure than those proposed in the preceding prospectus (ANOVA, MANOVA). Hierarchical linear modeling (HLM) will be presented in light of how it may be used within the context of an evaluation of the proposed professional development program. First, however, a description of some differences between traditional research and evaluation methods is presented in order to outline some of the considerations that have informed the proposed use of a HLM as a different statistical methodology.

Research versus Evaluation Methodologies

Experimental research, in the strictest sense, is characterized by a controlled experimental design and setting, and randomization of participants with the end goal being theory development and the ability to make claims about the causal nature of the relationship between variables under consideration. Evaluators often lack the same level of control over the program design—instead they collect data on an existing program to inform decision making processes regarding the program's effectiveness (Borich & Jemelka, 1981).

Besides the general feature and goal differences between evaluation and research, each method's approach to random selection and assignment may be different. Random assignment of participants to treatment groups is applied to control for factors that could represent competing alternatives and to assure that groups are equivalent at the outset of

the research (Shadish, Cook, & Campbell, 2002; Thompson, Diamond, McWilliam, Snyder, & Snyder, 2005). In evaluation, it is often administratively difficult to randomly assign participants to a group resulting in the use of quasi-experimental designs. While quasi-experimental designs have been defended, especially in evaluation settings (Campbell & Stanley, 1966 *cf* Shadish, Cook, & Leviton, 1991), the use of these methods do not allow for causal conclusions to be made about the relationship between treatment and outcome variables (Shadish, Cook, & Campbell, 2002). This contention is the case even when advanced statistical procedures are used (e.g., HLM) and follows the reasoning of esteemed methodologists in asserting that statistical analyses, no matter how sophisticated, cannot allow for implications regarding causality to be drawn—in other words, the experimental design, not the statistical tools, determine what types of implications can be drawn from research (Shadish, Cook, & Leviton, 1991).

In addition to participant selection, research and evaluation sometimes differ on the extent to which conclusions of a particular study may be generalized to other samples or groups (Borich, 2007b). Generalizability is ensured by selecting a sample that is reflective of a larger population. In research, attempts are made to select a sample that is reflective of a larger group of people, but evaluators do not always consider this a necessity, especially when they are focused on how a program operates within a specific context (Borich, 2007a). Since evaluation is concerned with a program's success in a particular context, an evaluator may not feel that it is necessary to form a sample representative of a larger group or context.

In closing, when engaged in an evaluation, it is essential to be aware of the differences between research and evaluation. For the particular purpose of this addendum, the primary issue faced by an evaluator would be the use of an alternate statistical procedure to account for features of the experimental design in an evaluation setting. In this case, the statistical procedures proposed by the preceding prospectus would not be ideal in the proposed evaluation context. Subsequent sections provide a more in-depth description of the program evaluation framework and outline the particular evaluation context posed in this addendum.

Evaluation Concepts within the Proposed Context

Varying approaches to evaluation are existent depending on evaluator characteristics, the evaluation setting, and the objectives of the evaluation (Borich & Jemelka, 1981). For the purposes of the current discussion, several hypothetical parameters are set in order to present a realistic evaluation situation as one in which HLM would be chosen as an analysis method. These parameters deal specifically with the role of the evaluator.

Among the roles an evaluator may adopt, most deal with the level of involvement the evaluator assumes in planning, designing, and/or evaluating a program. One crucial feature of an evaluator's role is to specify whether he or she is formatively or summatively involved with the program. Formative evaluations occur during the program by collecting data that help inform whether ongoing modifications are necessary. Summative evaluations are conducted at the program's completion to

determine whether it should continue (Borich, 2007b). This addendum presents the evaluation from the summative perspective where the evaluator would analyze data collected at the end of the program. In this sense, he or she would not affect change to the program as it proceeded.

Another important feature of the evaluator's role is whether he or she adopts a researcher, technician, decision-maker, or statistician perspective. For the purposes of the current discussion, the evaluator would adopt a statistician role implying that he or she would analyze data resulting from the program's data collection procedures. This role is in contrast with other perspectives such as the researcher or decision-maker, both of whom have more control over aspects of the program design, and not just the data analyses performed (Borich, 2007b). These tenets allow for a reconstruction of the original study to frame the argument for the use of HLM as a statistical tool.

Reconstruction of the Prospectus to Reflect Program Evaluation Perspectives

The prospectus research was presented from the perspective of a researcher with complete control over every aspect of the planning, procedure, and analyses. In this addendum, the assumption is that a program evaluator would not have the same control. In this hypothetical scenario, the program evaluator would be hired by a school district to assist with analyzing and reporting on data gathered after the previously described professional development exercise.

In the evaluation context, aspects of the original study would remain the same: the activities would still be designed to provide teachers with information on student assessment data and how these data could inform their teaching and assessment practices

and efficacy. The teachers would still be assigned to one of three groups (Personal Teaching Competence group, Task Analysis group, and Control group). Last, the professional development program would be provided in the summer and would extend slightly into the school year.

Several aspects of the program may change in an applied setting. In addition to the overall goal of the study shifting from theory development to an evaluation of the effectiveness of the professional development program, constraints present in realistic settings could prevent random assignment of teachers to groups. It is also likely that teacher selection would not be based on the proposed strata. According to Clement and Vandenberghe (2000), teachers are more likely to engage in professional development activities when they feel autonomous in choosing to participate in the activities. As it applies to the proposed program evaluation, this implies that it is more plausible that teachers would self-select into the exercise based on their personal preferences and availability. It is also more likely that they would volunteer for the professional development program rather than being selected by their principals.

Research and Program Evaluation Considerations Specific to the Prospectus

This prospectus proposed to develop the concept of teacher efficacy through an intervention-based research design. The intent was to determine whether teachers' sense of efficacy regarding their ability to utilize student assessment results would be amenable to change as a result of a three-week professional development intervention. As previously mentioned, the aim of this addendum is to refocus the description of the

prospectus methodology through the lens of program evaluation. It is important to reiterate that the procedures discussed below reflect this shift of focus—they present an alternative statistical approach to the data as if an evaluator was operating in a program evaluation context with little or no control over the experimental procedures. Therefore, this discussion follows as if the evaluator were taking on the role of a statistician analyzing on existent data in order to present findings to important stakeholders. Three areas of the original prospectus design are the focus of this revision: effects of the sampling method, benefits to the procedure, and HLM statistical analyses.

Experimental Design of Prospectus and Program Evaluation Considerations

Participant Selection: Prospectus versus Program Evaluation

To compile the sample, the prospectus proposed to use a stratified random sampling process with two strata: school performance ratings resulting from aggregate measures of student performance on standardized tests and teachers' level of experience. These strata were chosen based on prior research reporting their effects on the outcome variable under consideration (i.e., teachers' sense of efficacy) (e.g., Monasaas & Endelhard, 1994). In other words, the prospectus used this sampling procedure to experimentally control for the effects of previously-documented variables that may impact teachers' efficacy toward assessment. Due to limitations presented by the use of the school performance strata, only six elementary schools from each level of performance (e.g., Academically Unacceptable) were able to be chosen, resulting in four groups from six schools for a total of 24 schools. The second strata, teacher experience,

further limited the sample. The study proposed that principals from the 24 schools would be contacted to aid in the enlistment of teachers for the study. The principals would be asked to recruit eight teachers based on their level of experience: four novice teachers and four experienced teachers. In all, the resulting sample would include 192 elementary school teachers from the 24 selected schools.

From the perspective of a program evaluation, the sampling procedures described above may be unrealistic. For instance, it is unlikely that a sample chosen on the strata of performance rating and teacher experience would be possible—even if some of the participants were chosen according to these criteria, it would be difficult to gather a balanced sample. It is more plausible that this professional development program would have been offered to all teachers in the district without the advantage of teacher- and school-level controls through experimental design procedures. In this case, hierarchical linear modeling (HLM) would present an attractive alternative method to control for the effects of these teacher- and school-related characteristics.

Hierarchical linear modeling (HLM), also referred to as multilevel modeling, is a statistical procedure that allows for the examination of both fixed and random effects in hierarchically nested data structures such as teachers nested within classrooms and/or schools (Raudenbush & Bryk, 2002). The use of HLM would be beneficial in the context of proposed program evaluation due to its ability to account for teacher- and school-level factors. An evaluator would be able to control for the effects of teacher experience level and school performance rating through the analyses by adding these variables as random factors into the models at the teacher- and school-levels (level-1 and level-2). The

following regression equations represent the analyses that may occur for a research question where the outcome variable (Y_{ij}) is teachers' efficacy toward assessment, the predictor variable is group membership, and teacher experience level, pretest scores, and school performance rating represent covariates.

Level-1 (Teacher-level): $Y_{ij} = \beta_{0j} + \beta_{1j}(Group)_{1j} + \beta_{2j}(TeacherExperienceLevel)_{2j} + \beta_{3j}(PretestEfficacyScores)_{3j} + r_{ij}$

Level-2 (School-level): $\beta_{0j} = \gamma_{00} + \gamma_{01}(SchoolPerformanceRating)_j + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11}(SchoolPerformanceRating)_j + u_{1j}$

As a result of the statistical controls added to different levels of the data structure, an evaluator would not have to be as concerned about the lack of a stratified random sample because he or she could control for the effects of these variables statistically. This feature of the participant selection process would also rule out another problem associated with the experimental design of the prospectus in that it would not require that the principal select participants. In other words, eliminating this feature of the original study could also eliminate a potential source of bias in sample selection. In addition to the positive effects using HLM could have on how participants' characteristics could be controlled for statistically, HLM has implications for the experimental procedure.

Procedure: Prospectus versus Program Evaluation

In the proposal, participating teachers would be randomly assigned to one of three groups. Given a total sample size of 192 teachers, there would be 64 teachers in each of these groups, with equal representation of novice and experienced teachers from schools with varying levels of performance. The three-week long professional development

exercises would be slated to begin when TAKS scores (i.e., large-scale assessment scores) are released to schools in the summer and to extend slightly into the beginning of the school year.

One important limitation to this procedure is the assumption that teachers would continue to participate after the beginning of the school year. In the research-model proposed by the prospectus, this assumption is acceptable, but within the context of a real-world professional development, the evaluator could expect a certain level of participant attrition. Fortunately, HLM handles the effects of attrition—unequal cell sizes—better than analysis of variance procedures (Raudenbush & Bryk, 2002). This could be particularly fortunate since the research extends into the beginning of the school year, which is a busy and hectic time for teachers.

Proposed Analyses: Prospectus versus Program Evaluation

For the substantive research questions, analysis of variance techniques were slated for use. The first question asked whether posttest scores on the measurement concepts would be related to post-intervention scores on the teaching efficacy toward assessment survey. These analyses included a repeated-measures ANOVA (3 X 2) with one between-factor with three levels (group assignment) and one within-factor with two levels (measurement of subjects' assessment knowledge). The dependent variable was teachers' posttest performance on the assessment measure.

The second research question proposed to use a repeated-measures MANOVA. These analyses would have three dependent variables: task analysis, personal teaching competence and personal teaching efficacy. This analysis would have two between-

factors (group assignment and teacher experience) and one within-factor (posttest scores on teacher efficacy). In subsequent paragraphs, a discussion of how HLM would provide a much more powerful analysis tool is presented following a brief discussion of how these analyses proceed.

HLM analyses generally progress in two phases: the first stage consists of an analysis of the *unconditional* model. The model at the first stage does not include any predictors at the various levels of the nesting structure (e.g., teacher- or school-levels). The unconditional model is used to diagnose whether random variance in the intercepts and slopes exist at any level in the data hierarchy (Raudenbush & Bryk, 2002). In the second phase, predictor variables are added in order to test the *conditional model*. In terms of the proposed evaluation, the advantage of using HLM to analyze data are several: HLM would allow for the addition of more random effects than ANOVA, HLM would provide more statistical power, and HLM allows for the testing of cross-level interactions.

In statistics, a factor is considered random if it represents only a sample from some larger population (Raudenbush & Bryk, 2002). For example, participants in a study may be regarded as random since they represent a sample from a larger population. The benefit of using HLM over traditional least squares regression techniques and analysis of variance is that HLM models allow for more randomly varying factors to be incorporated into the analyses. This feature of HLM has implications for the ability to generalize findings to other populations since the sample (teachers) can be considered random. In the case of the current evaluation, the use of HLM could improve the external validity of

the study so that the findings could be generalized beyond the current sample of teachers. Also, and perhaps more importantly in an evaluation context, it would allow an evaluator to correctly model the factors under consideration. For instance, it would be optimal in this situation to model school membership as random since the schools in this study represent a sample of a larger population of schools. The same could be said for modeling the teachers as random. If the evaluator were to use traditional techniques, he or she would only be able to model one of these factors as random. With HLM, both teachers and schools can be modeled as random factors.

In addition to the ability to examine additional random effects in HLM, statistical power may be improved by the use of these models in several ways. First, as discussed, HLM appropriately models random effects. If an evaluator were to include a random effect without modeling for it appropriately, as is often done in traditional analysis of variance, standard error estimates may be inflated. As a result of this, observed outcomes can be affected—in fact, it is possible that important hypotheses can be rejected with an analysis of variance but would not be disconfirmed by an HLM model (Raudenbush & Bryk, 2002). This could be powerful for the current program evaluation discussion. Suppose a researcher were to conduct an ANOVA on these data, and due to larger standard errors resulting from the inappropriate inclusion of several random factors, he or she may determine that the professional development did not improve teachers' assessment skills or efficacy. However, if HLM was used by an evaluator, he or she could model random factors appropriately, and improve the analyses' ability to detect treatment effects on the outcome variables. Another advantage HLM has for statistical

power is that it allows for the addition of continuous and categorical independent variables rather than just categorical variables as is the case with ANOVA and MANOVA. This allows for a more fine-grained analysis of the effects of the independent variables on the dependent variable, which improves the power of the statistical test to detect differences that may occur at a wider array of the measurement continuum. Further, it would allow for a broader sample to be used—for instance, instead of qualifying teachers a priori based on their experience level (novice or experienced), information on teachers' tenure in education, which is typically measured continuously in years, could be used as a teacher-level covariate, and not as variable limiting sample selection. A final benefit of using HLM to improve statistical power includes analyzing variance components rather than just mean differences. By modeling variance at different levels of the data hierarchy, dependencies within groups may be explained (Raudenbush & Bryk, 2002). This allows the evaluator to adjust for these dependencies thereby more accurately estimating the standard error and inflating the value of the test statistic (e.g., *F* statistic). This would decrease the Type I error rate and the tendency to fail to reject the null hypothesis that the intervention does not impact post-program scores on the outcome measures.

The final advantage of using HLM in the currently discussed program evaluation context is that it allows for the examination of cross-level interactions. Cross-level interactions are particularly important in program evaluation because they reveal the person-environment interaction that is so crucial to determining whether a program would be effective within a given context. In other words, an existing cross-level

interaction may be uncovered using HLM, but would not be evident through the use of analysis of variance techniques.

Closing Remarks

The preceding discussion has presented logic for the use of HLM over traditional analysis of variance techniques in a hypothetical program evaluation scenario. This method offers a powerful alternative for statistical controls when experimental control was less plausible. Of course, as has been discussed, these statistical controls do not make up for a lack of random assignment or selection of participants, but they do allow evaluators to conduct more finely-tuned analyses on existent data.

Appendix A – The Collective Teacher Efficacy Scale (Goddard et al., 2007)

1. Teachers in this school have what it takes to get the children to learn.
2. Teachers in this school are able to get through to difficult students.
3. If a child doesn't learn something the first time, teachers will try another way.
4. Teachers here are confident they will be able to motivate their students.
5. Teachers in this school really believe every child can learn.
6. If a child doesn't want to learn teachers here give up.
7. Teachers here need more training to know how to deal with these students.
8. Teachers in this school think there are some students that no one can reach.
9. Teachers here don't have the skills needed to produce meaningful student learning.
10. Teachers here fail to reach some students because of poor teaching methods.
11. These students come to school ready to learn.
12. Homelife provides so many advantages they are bound to learn.
13. The lack of instructional materials and supplies makes teaching very difficult.
14. Students here just aren't motivated to learn.
15. The quality of school facilities here really facilitates the teaching and learning process.
16. The opportunities in this community help ensure that these students will learn.
17. Teachers here are well prepared to teach the subjects they are assigned to teach.
18. Teachers in this school are skilled in various methods of teaching.
19. Learning is more difficult at this school because students are worried about their safety.

20. Drug and alcohol abuse in the community make learning difficult for students here.
21. Teachers in this school do not have the skills to deal with student disciplinary problems.

APPENDIX B – Personal Teaching Efficacy Subscale of the TES

1. When a student does better than usual many times it will be because I exerted a little effort.

2. The hours in my class will have little influence of students compared to the influence of the home environment.

3. The amount that a student can learn in primarily related to family background.

4. If students aren't disciplined at home, they aren't likely to accept any discipline.

6. When a student gets a better grade than he/she usually gets, it will usually be because I found a better way of teaching that student.

7. When I really try, I will be able to get through to most difficult students.

8. A teacher is very limited in what he/she can achieve because a student's home environment is a large influence on his/her achievements.

9. When the grades of my students improve, it will usually be because I found more effective teaching approaches.

10. If a student masters a new concept quickly, this might be because I will have known the necessary steps in teaching that concept.

11. If parents would do more with their children, I could do more.

13. If a student in my class becomes disruptive and noisy, I feel assured that I will know some techniques to redirect him/her quickly.

15. If one of my students couldn't do a class assignment, I will be able to accurately assess whether the assignment was at the correct level of difficulty.

17. When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his/her home environment (RAND Item #1).

18. If I really try hard, I can get through to even the most difficult or unmotivated students (RAND Item #2).

Appendix C – Proposed Pre-Test of Measurement Concepts

ASSESSMENT CONCEPTS PRE QUIZ

(1) The raw score does not present a broad picture of test performance because it (Answer = A):

a. Can be interpreted only in terms of a particular set of test questions
b. Is not as interpretable as the percentage correct score
c. Is not a sufficient statistic for the model used in scoring the TAKS test
d. Must be turned into a scaled score to provide information about the student's performance

(2) Consider the following scenario: One year, 70% of students earned a raw score of 34 on a portion of the TAKS test. The following year, 75% of students earned a raw score of 34. The questions on the test in the second year were slightly easier than those on the test in the first year. What conclusion can you draw from the above scenario (Answer = C)?

a. There is no difference in the performance of students from the first to the second year
b. Students taking the test in the second year did better than those taking the test the first year
c. A conclusion cannot be drawn from the information provided
d. The improvement was due to chance fluctuation in scores from year to year

(3) Percentile rank scores represent (Answer = B):

a. Criterion-referenced scores
b. The percentage of students in the reference group earning scores below the score obtained
c. Scores equivalent to percentage correct scores
d. Scores that rank students from 1 to 100

(4) If your students are tested at a different time of the year than the norm group was tested, the interpretation of the percentile score is (Answer = B):

a. Valid
b. Unclear
c. Ok, as long as the students took the same version of the test
d. None of the above

(5) If a fifth grade student's math grade equivalent score is an 8.5, he (Answer = B):

a. Should be taking eighth grade math classes
b. Got as many right answers correct as an eighth grade student would have gotten if he had taken the fifth grade test
c. Performed better than 85% of the students in his class
d. Should be taking the eighth grade TAKS test in math

(6) Which is the greatest benefit to grade equivalent scores (Answer = D)?

a. They are usually properly interpreted
b. Unlike other types of scores, they have high accuracy for students who have very high scores
c. They can be used for computing group statistics
d. They are expressed in grade-level values that are familiar to parents

(7) The TAKS test reports students' scores in scale score format. A scale score is *not* a score that (Answer=D):

a. Has been converted onto a scale common to all test forms for that assessment
b. Takes into account the difficulty level of the specific set of questions
c. Relates information about a student's performance relative to passing standards
d. Relates the number of items correctly answered in a section

(8) Using the provided formula and scale score table for converting a scale score to a percentile rank. Choose the correct answer among the options below (Answer = C).

a. 94
b. 82
c. 89
d. None of the above

(9) Which of the following pieces of information is *best* represented by letter C on the score report (Answer = C)?:

a. Scale score
b. Raw score
c. Written Composite Rating
d. Lexile

(10) In which of the tested areas does this student need improvement in order to meet the standard (Answer = A)?

a. English/Language Arts
b. Mathematics
c. Social Studies
d. Science

Appendix D – Proposed Posttest of Measurement Concepts

ASSESSMENT CONCEPTS POST QUIZ

(1) The raw score does not present a broad picture of test performance because it (Answer = B):

a. Is not a sufficient statistic for the model used in scoring the TAKS test
b. Can be interpreted only in terms of a particular set of test questions
c. Must be turned into a scaled score to provide information about the student's performance
d. Is not as interpretable as the percentage correct score

(2) Consider the following scenario: One year, 80% of students earned a raw score of 29 on a portion of the TAKS test. The following year, 85% of students earned a raw score of 29. The questions on the test in the second year were slightly easier than those on the test in the first year. What conclusion can you draw from the above scenario (Answer = A)?

a. A conclusion cannot be drawn from the information provided
b. Students taking the test in the second year did better than those taking the test the first year
c. The improvement was due to chance fluctuation in scores from year to year
d. There is no difference in the performance of students from the first to the second year

(3) Percentile rank scores represent (Answer = C):

a. Scores that rank students from 1 to 100
b. Scores equivalent to percentage correct scores
c. The percentage of students in the reference group earning scores below the score obtained
d. Criterion-referenced scores

(4) If your students are tested at a different time of the year than the norm group was tested, the interpretation of the percentile score is (Answer = C):

a. Valid
b. Ok, as long as the students took the same version of the test
c. Unclear
d. None of the above

(5) If a fourth grade student's math grade equivalent score is an 6.5, he (Answer = D):

a. Should be taking sixth grade math classes
b. Performed better than 65% of the students in his class
c. Should be taking the sixth grade TAKS test in math
d. Got as many right answers correct as an sixth grade student would have gotten if he had taken the fourth grade test

(6) Which is the *greatest* benefit to grade equivalent scores (Answer = A)?

a. They are expressed in grade-level values that are familiar to parents
b. Unlike other types of scores, they have high accuracy for students who have very high scores
c. They can be used for computing group statistics
d. They are usually properly interpreted

(7) The TAKS test reports students' scores in scale score format. A scale score is *not* a score that (Answer=B):

a. Relates information about a student's performance relative to passing standards
b. Relates the number of items correctly answered in a section
c. Takes into account the difficulty level of the specific set of questions
d. Has been converted onto a scale common to all test forms for that assessment

(8) Using the provided formula and scale score table for converting a scale score to a percentile rank. Choose the correct answer among the options below (Answer = C).

a. 94
b. 82
c. 89
d. None of the above

9) Which of the following pieces of information is *best* represented by letter H on the score report (Answer = D)?:

a. Scale score
b. Raw score
c. Written Composite Rating
d. Lexile Measure

(10) In which of the tested areas does this student need improvement in order to meet the standard (Answer = A)?

a. English/Language Arts
b. Mathematics
c. Social Studies
d. Science

Appendix E – Example of TAKS Score Report Used in the Proposed Study

<Insert Sample Score Report Here>

REFERENCES

American Federation of Teachers, National Council on Measurement in Education and National Education Association. (1990). *Standards for Teacher Competence in Educational Assessment of Students*. Washington, DC: American Federation of Teachers.

Armstrong, J., & Anthes, K. (2001). How data can help. *American School Board Journal, 188*, 38–41.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.

Bandura, A. (1989). Social cognitive theory. In R. Vasta (Ed.), *Annals of Child Development (Vol.6): Six Theories of Child Development* (pp. 1-60). Greenwich, CT: JAI Press.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117-148.

Bernhardt, V. L. (2000). Intersections. *Journal of Staff Development, 21*(1), 33–36.

Betz, N. E. (2000). Self-efficacy theory as a basis for career assessment. *Journal of Career Assessment, 8*, 205-222.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-68.

Borich, G. (2007a). *Some Distinctions between the Researcher and the Evaluator* (class handout). Evaluation Models and Techniques, Spring 2007, The University of Texas at Austin.

Borich, G. D. (2007b). *Program Evaluation: Models and techniques (course packet)*. Evaluation Models and Techniques, Spring 2007, The University of Texas at Austin.

Borich, G., & Jemelka, R. (1981). Definitions of program evaluation and their relation to instructional design. *Educational Technology, 21*(8), 31-38.

Brown, G. T. L. (2002). *Teachers' Conceptions of Assessment.* Unpublished Doctoral Dissertation, University of Auckland, Auckland, NZ.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., & Wexler, D. (2005). Linking data and learning: The grow network study. *Journal of Education for Students Placed at Risk, 10,* 241-267.

Chen, J. Q., Salahuddin, R., Horsch, P., & Wagner, S. L. (2000). Turning standardized test scores into a tool for improving teaching and learning: An assessment-based approach. *Urban Education, 35*, 356-384.

Clement, M., & Vandenberghe, V. (2000). Teachers' professional development: A solitary or collegial (ad)venture? *Teacher and Teaching Education, 16*, 81-101.

Codding, R. S., Skowron, J., & Pace, G. M. (2005). Back to basics: Training teachers to interpret curriculum-based measurement data and create observable and measurable objectives. *Behavioral Interventions, 20*, 165-176.

Confrey, J., & Makar, K. (2002). Comparing two distributions: Investigating secondary teachers' statistical thinking. *Statistics Education Research Journal, 1*(1), 38-40.

Confrey, J., Makar, K. M., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *International Reviews on Mathematics Education (ZDM), 36*(1), 32-40.

Confrey, J., & Makar, K. (2005). Critiquing and improving data use from high stakes tests: Understanding variation and distribution in relation to equity using dynamic statistics software. In C. Dede, J. P. Honan, & C. Peters (Eds.), *Scaling up Success: Lessons Learned from Technology-Based Educational Improvement* (pp. 198-226). San Francisco: Jossey-Bass.

Creighton, T. B. (2001). *Schools and Data: The Educator's Guide for Using Data to Improve Decision Making*. Thousand Oaks, Calif: Corwin Press.

Cromey, A. (2000). *Using assessment data: What can we learn from schools*? Naperville, IL: North Central Regional Education Laboratory.

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

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students*. Center on Educational Governance: University of Southern California.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal, 44*, 594 –629.

Foley, D. W. (2007). *Building Capacity among Elementary Teachers Using Data*. Unpublished Doctoral Dissertation, University of Pittsburgh, Pittsburgh, PA.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76,* 569-582.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment.* London: The Falmer Press.

Goddard, R. D., Hoy, W. K., & Woolfolk Hoy, A. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, *37*, 479–507.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through preservice teacher education. *Teaching and Teacher Education, 21*, 607–621.

Green, K. E., & Stager, S. F. (1986). Measuring teacher attitudes toward testing. *Measurement and Evaluation in Counseling and Development, 19,* 141-150.

Guskey, T. R., & Passaro, P. D. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal, 31*, 627-643.

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26*(1), 19-27.

Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education, 74,* 262–273.

Henson, R. K. (2001). The effects of participation in teacher research on teacher efficacy. *Teaching and Teacher Education, 17*, 819-836.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

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

Henson, R. K., Bennett, D. T., Sienty, S. F., & Chambers, S. M. (2002). The relationship between means-end task analysis and context-specific and global self-efficacy in emergency certification teachers: Exploring a new model of self-efficacy. *The Professional Educator, 24*(2), 29–50.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Teachers' ability to interpret standardized test scores. *Educational Measurement: Issues and Practice, 10*(4), 16-18.

Impara, J.C. & Plake, B.S. (1996). Professional development in student assessment for educational administrators: An Instructional Framework. *Educational Measurement: Issues & Practice, 15*(2), 14-19.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*, 1258-1287.

Kazak, S. & Confrey, J. (2004). *Investigating educational practitioners' statistical reasoning in analysis of student outcome data.* Paper presented by distribution at the 10th International Congress on Mathematical Education (ICME), Copenhagen, Denmark.

Keith, T. Z. (2006). *Multiple Regression and Beyond.* Boston: Allyn & Bacon.

Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*, 496-520.

Klein, T. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. Thousand Oaks, CA: Sage Publications.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T. & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81,* 199–203.

Labone, E. (2004). Teacher efficacy: Maturing the construct through research in alternative paradigms. *Teaching and Teacher Education, 20*, 341-359.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk, 10*, 333-349.

Louis, K. S., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis, 27*, 177–204.

Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision making. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 13-31). New York: Teachers College Press.

Mandinach, E. B., Rivas, L., Light, D., Heinze, C. & Honey, M. (April, 2006). *The impact of data-driven decision making tools on educational practice: A systems analysis of six school districts*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Marso, R. N., & Pigge, F. L. (1988). Ohio secondary teachers testing needs and proficiencies: Assessments by teachers, supervisors, and principals. *American Secondary Education, 17*, 2-9.

Mokhtari, K., Rosemary, C. A., & Edwards, P. A. (2007). Making instructional decisions based on data: What, how, and why. *The Reading Teacher*, *61*, 354-359.

Monasaas, J. A., & Endelhard, G. (1994). Teachers' attitudes toward testing practices. *The Journal of Psychology, 128*, 469-477.

Mulvenon, S., Stegman, C., & Ritter, G. (2005). Test anxiety: A multifaceted study on the perceptions of teachers, principals, counselors, students, and parents. *International Journal of Testing, 5*(1), 37-61.

Protheroe, N. (2001). Improving teaching and learning with data-based decisions: Asking the right questions and acting on the answers. *ERS Spectrum 19*(3), 4–9.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.).* Thousand Oaks, CA: Sage Publications.

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. *Teaching and Teacher Education, 10*, 381-394

Ross, J. A. & Bruce, C. (2007a). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education, 23,* 146–159.

Ross, J. A. & Bruce, C. (2007b). Professional development effects on teacher efficacy: Results of randomized field trial. *Journal of Educational Research, 101*(1), 50-60

Rudner, L. & W. Schafer (2002) *What Teachers Need to Know About Assessment*. Washington, DC: National Education Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, CA: Sage Publications.

Stein, M. K., & Wang, M. C. (1988). Teacher development and school improvement: The process of teacher change. *Teaching and Teacher Education*, *4*(1), 171-187.

Stevens, J. P. (1999). *Intermediate Statistics: A Modern Approach* (Second Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Supovitz, J. A., & Klein, V. (2003). *Mapping a Course for Improved Student Learning: How Innovative Schools Systematically Use Student Performance to Guide Improvement.* Consortium for Policy Research in Education (University of Pennsylvania). Retrieved May 10, 2008 from: http://www.cpre.org/images/stories/cpre_pdfs/AC-08.pdf

Texas Assessment of Knowledge and Skills (TAKS) Usage Manual (2008). Austin, TX: Texas Education Agency. Retrieved January 24, 2008 from: http://www.tea.state.tx.us/student.assessment/teachers.html

Thompson, B., Diamond, K. E., McWilliam, R., Synder, P., & Synder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71,* 181-194.

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68,* 202-248.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17,* 783–805.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522.

Wayman, J. C. (2005). Involving teachers in data-based decision-making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed At Risk, 10,* 295-308.

Wayman, J. C. & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education, 112,* 549-571.

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). *The data-informed district: A district-wide evaluation of data use in the Natrona County School District.* Austin: The University of Texas.

Williams, J., & Ryan, J. (2000). National testing and the improvement of classroom teaching: Can they coexist? *British Educational Research Journal, 26*, 49-73.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37–42.

Wolfe, E. W., Viger, S. G., Jarvinen, D. W., & Linksman, J. (2007).Validation of scores from a measure of teachers' efficacy toward standards-aligned classroom assessment. *Educational & Psychological Measurement, 67*, 460-474.

Woolfolk Hoy, A., & Burke-Spero, R. (2005). Changes in teacher efficacy during the early years of teaching: A Comparison of four measures. *Teaching and Teacher Education, 21*, 343-356.

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

VITA

Shana Michele Shaw was born in Lake Forest, Illinois on June 5, 1979, the daughter of Debra Joan Margoliner and James Dean Shaw. After completing her work at Angleton High School, Angleton, Texas, in 1997, she entered Stephen F. Austin State University in Nacogdoches, Texas. She received the degree of Bachelor of Arts from Stephen F Austin State University in August, 2001. She also earned a Masters of Education from Texas Tech University in May, 2005. In September, 2005, she entered the Graduate School at the University of Texas at Austin.

Permanent Address: 3100 Fontana Drive

Austin, Texas 78704

This report was typed by the author