Teachers' Perceptions of the Use of Student Learning Measures in Teacher Evaluation: An Examination of the Use of Student Growth Percentiles in Virginia

---

A Dissertation
Presented to the
Faculty of the Curry School of Education
University of Virginia

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

By
Michael C. Irani, A.B., M.A.

ABSTRACT

Dr. Pamela D. Tucker

Current educational reform for public K-12 schools has focused on revising teacher evaluation policies and practices. One of the chief revisions has been an emphasis on student learning measures as part of a teacher's summative evaluation. This study examines the perceptions of teachers in Virginia in regards to using one type of student learning measure, student growth percentiles, in a teacher's evaluation. The study gathered teachers' attitudes toward the practice as it relates to the areas of propriety, utility, and accuracy, three of the four standards that are delineated in *The Personnel Evaluation Standards* (Gullickson, 2009). A survey based on these three standards was designed and both closed and open ended responses were collected from 150 teachers in the Commonwealth.

Results of the study indicate that teachers have a negative attitude toward the use of student growth percentiles in a teacher's evaluation in all three domains. Teachers were particularly concerned about the accuracy of using student growth percentiles in a teacher's evaluation primarily because they feel the practice does not allow for the accounting of outside influences on students' achievement on standardized tests among other factors. Teachers also expressed considerable concern about the potential negative impact on collegiality in a school that used student growth percentiles. These broad findings along with others are discussed in this study and are used to provide recommendations for practice and further research. Recommendations for further practice include improving the accuracy and

fairness of the practice, communicating how student growth percentiles are used more clearly, and considering the possible impact of student learning measures on a school's culture. Recommendations for further research include increasing the sample size of participants, analyzing how other types of student learning measures are perceived, and exploring specific ways in which student growth percentiles are used in the supervisory process with teachers.

SIGNATURE PAGE

ACKNOWLEDGEMENT

I would like to thank the members of my dissertation committee for their guidance and support throughout the process of completing this study. I would specifically like to thank Dr. Pamela Tucker for her expertise in the field and commitment to being a teacher, Dr. Patrick Meyer for his patience and analytical mind, and Dr. Denny Berry for her ability to keep the study grounded in the current educational context and policies.

I have also been fortunate enough to work in two great school divisions as I have completed my doctoral work. I would like to especially thank Drs. Pamela Moran and Rosa Atkins, Mrs. DeeDee Jones, and Dr. Thomas Taylor for their support, encouragement, and understanding during my studies.

Finally, I have a family second to none and owe a tremendous amount of this to them. To my wife, Stephanie, and two children, Jonathan and Maddie, thank you for putting up with me – I look forward to having Sunday mornings and school holidays back with you!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Background

Recent international and national reports have painted a troubling picture of the state of American K-12 public education. In terms of literacy, the Programme for International Student Assessment (PISA) ranked the United States 14th out of 34 countries (Associated Press, December 7, 2010). The National Assessment of Educational Progress (NAEP) revealed that 4th grade reading scores had gone unchanged between 2009 and 2011, despite numerous reform efforts (National Assessment of Educational Progress, 2011). Similarly, there was no change for 8th grade students at our above the basic level of literacy achievement between 2009 and 2011 (NAEP, 2011). Math achievement of American students has been even more disappointing to public officials as the United States ranked 25th out of 34 countries on the PISA examination (Associated Press, December 2, 2010) and the World Economic Forum ranked the US 48th in math education (Klein, 2011). PISA also noted that of the 34 countries assessed in its study that only 8 countries had a lower graduation rate (Associated Press, December 7, 2010); Klein (2011) cites the graduation rate itself to be hovering around 70%. Even when students do graduate high school, the ACT noted that only 24% of high school graduates were adequately prepared for entry level college courses (ACT, 2010).

Such reports of the state of American education have coincided with strong political rhetoric and academic work centered on the role of the teacher and teacher evaluation. For example, in light of a report detailing the impact of teacher layoffs, President Obama stated:

> If we want America to lead in the 21st century, nothing is more important than giving everyone the best education possible — from the day they start preschool to the day they start their career. (The White House, August 18, 2012)

A major focus in reforming U.S. education has been on teacher quality. Former secretary of education, William Bennett, recently cited a study claiming that it shows that "second only to parents, teachers are the most important part of a child's education" (Bennett, 2012, para. 13). As noted by Marzano (2012), however, several studies have highlighted how traditional teacher evaluation systems have failed to identify effective or ineffective teachers, a particularly troubling conclusion considering the political focus on providing a high quality education to children.

With the political and academic focus on using teacher evaluation as one method in identifying effective teachers and providing a high quality education, it is important to understand how past methods have fallen short. Danielson and McGreal (2000) argued that high quality teacher evaluation has occurred, at best, rarely in the past. In addition, teacher evaluation systems are not viewed as helpful by teachers themselves (Teoh, 2012). As Stronge (2006) notes, "Too often, educational reform has produced disappointing results … A conceptually sound and properly implemented evaluation system for teachers is a vital component for successful reform efforts" (p. 3). Difficulty in creating a sound teacher evaluation system may be due, in part, to perception that teaching is a multi-faceted act, one that may be described as

delivering knowledge or motivating a learner to some (Kennedy, 2010) or even the nurturing of a student's social and personal development in addition to academic growth (Brophy, 1986). Regardless of the definition, it is generally accepted that an effective teacher evaluation system is critical to creating high quality schools (Cowart & Myton, 1997; Stronge, 1997). In addition to being effective, a teacher evaluation system must also be fair in order to foster growth among teachers and schools (Danielson & McGreal, 2000; Stronge, 1997; Stronge, Ward, Tucker, Hindman, McCloskey, & Howard, 2007). Balancing effectiveness and fairness, however, has been a traditional struggle for school administrators and policymakers (Marshall, 2009; Peterson, 2000; Ravitch, 2010; Stufflebeam, 1997).

Student Learning Measures in Teacher Evaluation

Among the aspects of teacher evaluation that has drawn the most attention in recent years in regards to striking this balance has been the use of student learning measures (Peterson & Peterson, 2006). The call for the inclusion of student learning as part of a balanced teacher evaluation system stems from dissatisfaction with older methods (Milanowski, 2011; Nolan & Hoover, 2004; Peterson, 2000; Peterson & Peterson, 2006; Stronge & Tucker, 2000; Stronge et al., 2007; Weisberg, Sexton, Mulhern, & Kaeeling, 2009). For example, evaluators often correlated advanced degrees and experience with high quality teaching but research has shown that these are weak indicators at best (Gallagher, 2004; Hanushek, 1971; Hanushek & Rivkin, 2007; Harris & Sass, 2009; Koedel & Betts, 2005; Munoz & Chang, 2007). In addition, a pivotal study indicated that a decade's worth of focus on using standards for teacher evaluation did not change the actual evaluation practices considerably in the U.S.'s 100 largest school districts (Loup, Garland, Ellett, & Rugutt, 1996). This finding may be largely related to others' criticisms

that principals lack the content knowledge or inter-rater reliability to provide a meaningful and accurate evaluation (Milanowski, 2011; Weisberg et al., 2009). Furthermore, it has been estimated that a principal is able to observe 0.1% of a teacher's actual teaching over the course of a year and of these observations, many may not be authentic representations of what occurs on a day-to-day basis (Marshall, 2005). In addition to the dissatisfaction with past evaluation practices, there is a growing desire to focus teacher evaluation on the core purpose of teaching, student learning (Cowart & Myton, 1997; Fenstermacher & Richardson, 2005; Hanuchek & Rivkin, 2007; McConney, Schalock, & Schalock, 1997; Peterson, 2000; Rivkin, Hanushek & Kain, 2005; Stronge, Ward, Tucker, & Hindman, 2007). Research has also strongly indicated that the single most important factor in accounting for student learning is the teacher (Rivkin, Hanushek & Kain, 2005; Sanders & Rivers, 1996; Stronge & Tucker, 2000; Wright, Horn, & Sanders, 1997). For example, Sanders and Rivers (1996) found that teachers rated in the top quintile facilitated adequate academic progress in all students while teachers rated in the lowest quintile made unsatisfactory gains regardless of their students' previous achievement on standardized tests. More recently, research has found that a teacher judged to be one-standard deviation above average can lead to dramatic achievement gains in both math and reading (Koedel & Betts, 2007). Hanushek and Rivkin (2010) have indicated that the difference in having a teacher in the 25th percentile compared to one in the 75th percentile are learning gains of 0.2 standard deviations, a sizable impact. Hanushek (2011) has recently made two compelling arguments regarding the impact of the teacher, concluding that if the U.S. were to replace its bottom 5-7% of least effective teachers with average teachers, the U.S.'s achievement would match that of countries such as Finland, which has been held up by many as a paragon for its

education system. Furthermore, Hanushek (2011) also estimates that a teacher in the 69th percentile produces students who earn $10,600 more in a lifetime than a teacher in the 50th percentile; similarly, a teacher in the 16th percentile produces students who make $400,000 less over the course of a lifetime when compared to an average teacher.

Current Educational Reform

The inclusion of student academic gains to teacher evaluation has become a centerpiece of current educational reforms (Newton, Darling-Hammond, Haertel, & Thomas, 2010). When the Bush administration enacted No Child Left Behind (NCLB), one of the calls was to ensure that a "highly qualified" teacher was in every classroom (Berliner, 2005; Phillips, 2010); the movement under the Obama administration has been to ensure that every child has a "highly effective" teacher (Darling-Hammond, 2009). In *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act* (2010), the United States Department of Education (USDOE) under the leadership of President Obama's Secretary of Education, Arne Duncan, identifies "focusing on teacher and leader effectiveness in improving student outcomes" (USDOE, 2010, p.13) as one of the core principles of the reauthorization. Specifically, one of the requirements is:

> Statewide definitions of "effective teacher," "effective principal," "highly effective teacher," and "highly effective principal," developed in collaboration with teachers, principals, and other stakeholders, that are based in significant part on student growth … (USDOE, 2010, p.14)

In addition, districts are to craft evaluation systems that implement each state's definitions of the different levels of "effectiveness" (USDOE, 2010). On September 23, 2011, Secretary Duncan

officially invited State Chief School Officers to apply for waivers to NCLB, encouraging them to use this refocused approach to teacher evaluation as part of their application (USDOE, 2011).

Virginia is one of the states that applied for and received a waiver from provisions in NCLB that stipulated that all students would be proficient in reading and mathematics by 2014. In the waiver approval letters, Secretary Duncan (2012) cites the application's commitment to revising "teacher evaluation … systems that support student achievement" as one of the key components of the request. Virginia had been revising its teacher evaluation system prior to the application, approving its *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers* on April 28, 2011. Chief among the revisions to these guidelines was a change to make "student academic progress" account for 40% of a teacher's summative evaluation (Virginia Department of Education, 2011b).

Districts in Virginia have since worked to make sense of the increased focus on "student academic progress" and how to tie it into their evaluation systems. The VDOE has stated that assessment of this domain be "determined by multiple measures of learning and achievement, including, when available and applicable, student-growth data from VDOE" (VDOE, 2011a, para. 2). One of the tools provided to school districts by the VDOE are student growth percentiles, defined as an expression of "how much progress a student has made relative to the progress of students whose achievement was similar on previous assessments" (VDOE, 2011d, para. 1). The VDOE continues to state:

> Student growth percentiles are calculated by identifying all students in the state whose previous SOL scaled scores in a subject are statistically similar and, then, comparing the achievement of these students on the next grade-level test. The achievement of each

student relative to that of the other students in the group is expressed as a percentile. (VDOE, 2011d, para. 3)

Because of the availability of tests given on an annual basis, the growth percentiles are available for students in Grades 4-9 in math and reading (VDOE, 2011d).

Since the student growth percentiles are provided as one way that districts may choose to measure student academic growth in a teacher's summative evaluation, it is unclear how many districts have actually used them. Virginia's Performance-Pay Initiative, however, requires schools that take part to use the percentiles (when applicable) as part of their teacher evaluation system (VDOE, 2011c). The Performance-Pay Initiative selected 25 schools in 13 districts to take part, providing them funds to pay teachers a bonus in 2011-2012 and 2012-2013 who earned "exemplary" ratings, using the revised teacher evaluation guidelines as a model (VDOE, 2011c).

Problem

Despite the growing momentum to use student learning measures in teacher evaluation in states such as Virginia, there has been a tremendous amount of concern expressed regarding the practice. Briefings to policy makers have warned that the use student test scores in and of themselves as a basis for making high stakes decisions is inappropriate (Baker, Barton, Darling-Hammond, Haertel, Linn, Ravitch, …, Shepard, 2010; Hinchey, 2010) and are a "crude indicator" of a teacher's contributions to students' learning (Corcoran, 2010). One of the more vocal and influential voices on the topic, Darling-Hammond (2009) asserts that extreme caution should be used when determining the specific role a teacher has on student learning because of the various other factors outside of a teacher's control that influence a student's achievement.

Teacher unions have echoed the concerns communicated to policy makers. The National

Education Association has stated:

> Unfortunately, the use of student learning measures to improve teaching practice has too often translated into using "value-added" analyses of state standardized test scores as the primary, or even sole, means for making summative decisions about teachers. Such use of test data is inappropriate for many reasons that are well documented. (NEA, 2010, p. 8)

Chief among the reasons for the NEA's stance are that a single test score does not accurately represent a student's learning, analysis of the data is largely dependent on the method used, the students assigned to teachers largely determine outcomes, outside factors can affect outcomes in ways one cannot measure, and most teachers teach subjects that cannot be measured in using standardized tests (NEA, 2010). The American Federation of Teachers has also voiced concerns regarding the use of student achievement data in teacher evaluation. In a position paper, the AFT states:

> Student learning should include evidence of growth in knowledge and skills based on multiple measures. Just as no one measure can evaluate teacher performance, no one measure can or should account for student learning. (AFT, 2010, p. 6)

The AFT focuses its criticisms more on the use of standardized tests, claiming that they should not be either the sole or primary measure of learning (AFT, 2010).

Teachers in Virginia have voiced specific concerns regarding the way the Virginia Department of Education has made use of student performance data in teacher evaluation. In response to the adoption of the teacher evaluation guidelines, the Virginia Education Association president outlined several concerns, including:

- a lack of time of resources to properly implement the new guidelines,

- the emphasis on standardized tests and their negative impact on schooling,
- the use Virginia SOL tests to create a growth measure when they may not be accurate enough, and
- technical concerns regarding the creation of the Student Growth Profile, especially as they relate to high achieving students. (VEA, 2011)

Despite the articulation of these concerns, the VDOE adopted the revised teacher evaluation guidelines and districts are currently implementing them.

Purpose and Research Questions

The purpose of this study was to determine teachers' perceptions of the use of student performance data in the teacher evaluation process with a focus on three key areas: the propriety, utility, and accuracy of the practice. These three areas were selected because they were the most relevant domains of The Professional Evaluation Standards (Gullickson, 2009) related to using student learning measures in teacher evaluation. In addition, little research has been conducted on teachers' perceptions of the topic. A focused study on the topic may have a dramatic effect on the work that district administrators and principals do when developing teacher evaluation systems and related professional development opportunities.

The following research questions guide this study:

1. What are teachers' perceptions of the use of student growth percentiles in teacher evaluation in terms of propriety?
2. What are teachers' perceptions of the use of student growth percentiles in terms of utility?

3. What are teachers' perceptions of the use of student growth percentiles in terms of accuracy?

Rationale

The rationale for this study is grounded in the fact that there has been little research conducted on teachers' perceptions of the use of student learning measures in teacher evaluation. The work that has been completed recently has focused on the general topic as part of a survey to measure teacher beliefs about various teacher policies (Donaldson, 2012; Teoh & Coggins, 2012). In addition, there has been no research published on teachers' perceptions of the use of student growth percentiles used in Virginia schools. A large body of work has been completed on related topics such as the accuracy of using student learning measures in teacher evaluation (Baker et al., 2010; Cantrell & Kane, 2012; Darling-Hammond, 2009; Hanuchek & Rivkin, 2010; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rockoff & Speroni, 2010; Stronge, Ward, & Grant, 2011), outside factors that influence achievement other than a teacher (Baker et al., 2010; Corcoran, 2010; Darling-Hammond & Rustique-Forrester, 2005; Kane & Steiger, 2008; Rivkin, Hanuchek, & Kain, 2005; Steele, Hamilton, & Stecher, 2010), stability of models used (Amrein-Beardsley, 2008; Lockwood, McCafferey, Hamilton, Stechner, Vi-Nhuan, & Martinez, 2007; Palardy, 2010), and the impact that the practice could have on instruction and school climate (Anderman, Anderman, Yough, & Gimbert, 2010; Cooper, Ehrensal, & Bromme, 2006; Marshall, 2005). Individual teachers' perceptions expand the body of literature on the relatively new practice of using student learning measures in evaluation and provide a perspective not yet fully explored in detail. In addition, the findings of the study will help inform the work of policy makers, district administrators, and principals. Specifically, the study

elucidates concerns that unions have voiced to policy makers. Furthermore, the study provides school and district administrators a better understanding of their teachers' beliefs regarding the use of student learning measures in evaluation, knowledge that will inform how they develop and implement teacher evaluation policies and procedures with a faculty as well as structure professional learning opportunities.

Overview of Study

This study analyzed the perceptions of teachers' beliefs about the use of student growth percentiles in teacher evaluation. An on-line survey was the central tool for data collection and asked closed and open-ended questions in order to gain both a wide scope of teachers' perceptions and a richer understanding of their beliefs. The survey was framed around the relevant standards delineated in *The Personnel Evaluation Standards* (Gullickson, 2009).

Limitations

A major limitation of the study is the small sample from which it drew. One-third of the sample were teachers that have had student growth percentiles applied in their evaluation as indicated by their school's participation in the Virginia Pay for Performance program. The remaining percentage of the sample are teachers in like schools who may or may not have had the student growth percentiles used in their evaluation and who may or may not have been eligible to receive additional pay for performance. As such, the results do not represent generalized conclusions of all teachers who have student growth percentiles used in their evaluation, rather just those taking part in the Virginia Pay for Performance program. In addition, the perceptions reported from teachers of like schools may not represent generalized conclusions the way a pure random sample would. The sample also focused on teachers of

reading and mathematics in grades 4-8, excluding the perceptions of teachers in other grades and content areas. Finally, the study was limited to Virginia teachers, making generalizations beyond the commonwealth impossible.

Definitions

The following terms are used throughout this study; for the purpose of clarification, their definitions are as follows:

1. Teacher evaluation – The systematic assessment of a teacher's performance in relation to his/her role and assignment in a school (adapted from Gullickson, 2009).
2. Student achievement – The representation of what a student has learned in terms of knowledge and skills, most often in the form of standardized tests (adapted from McConney et al., 1996).
3. Student growth – Demonstrated change in student achievement from one point in time to another (adapted from Betenbenner & Linn, 2009).
4. Student Growth Percentile – As defined by the Virginia Department of Education (2011):

> A student growth percentile describes how much progress a student has made relative to the progress of students whose achievement was similar on previous Standards of Learning (SOL) assessments in either reading or mathematics. Student growth percentiles are calculated by identifying all students in the state whose previous SOL scaled scores in a subject are statistically similar, and then comparing the achievement of these students on the test they take in the next

grade level. The performance of each student relative to that of the other students in the group is expressed as a percentile.

Organization of the Study

Chapter 2 of this study presents a review of the literature related to the use of student achievement data in teacher evaluation. The review is organized using key themes established by *The Personnel Evaluation Standards* (Gullickson, 2009), the resource used by most states and districts when assessing teacher evaluation policies and practices. Chapter 3 describes the methods, instruments, and procedures to be used in the study. Chapter 4 details the major findings of the study. Finally, Chapter 5 summarizes the findings and provides recommendation for further research and action steps to be taken by stakeholders.

CHAPTER 2: LITERATURE REIVEW

Purpose

The purpose of Chapter Two is to review the literature related to the use of student learning measures in the context of teacher evaluation. As a framework, the chapter will first provide an overview of different models used for measuring student learning in the context of teacher evaluation. The chapter will then examine the literature using the relevant standards from *The Personnel Evaluation Standards* (Gullickson, 2009). A description of each standard will be followed by a review of the literature associated with it.

Accountability Models for Measuring Student Learning

The methods used by states to measure student learning for the purposes of teacher evaluation vary considerably. When No Child Left Behind was first reauthorized by the Bush administration in 2001, the primary method was through the use of criterion-referenced assessments while the waivers awarded by the Obama administration have pushed the focus toward growth models (Carey & Manwaring, 2011). States continue to use methods that reflect the gamut of these two approaches. To frame the various approaches, the Council of Chief State School Officers (2005) commissioned a paper to detail these methods. The paper details four main models: the status model, the improvement model, the growth model, and the value-added model (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, & Williams, 2005).

Status Model

Goldschmidt et al. (2005) define a status model as one that measures the achievement levels of school as compared to an established target. In a status model, student achievement is classified in categories such as "advanced", "proficient", "below proficient", and "below basic" based on a score on a standardized test (The Center for Public Education, 2007). In some cases, a scaled score may be used where a raw score is converted to a scaled score according to the difficulty of the assessment (The Center for Public Education, 2007). The status model is one that had been used under NCLB as states measured the proficiency levels of sub-groups according to annual measurable objectives, targets established by individual states (Goldschmidt et al., 2005). Individual states have also elected to use status models to evaluate schools and school districts; for example, Virginia's school accreditation ratings establish separate targets for proficiency in order to earn different levels of accreditation by the state department of education (VDOE, 2012). The status model has been praised for its ability to shed light on what groups of students are not succeeding at desired levels, but has also been criticized for not focusing on the growth of the individual student (Carey & Manwaring, 2011).

Improvement Model

The improvement model is defined by Goldschmidt et al. (2005) as a status model that "measures change between different groups of students" (p. 3). Despite this definition, The Center for Public Education (2007), an initiative of the National School Boards Association, groups the improvement model under the umbrella of growth models because it has been allowed as one under the provisions laid out by NCLB. In addition, the performance index model has been considered a form of an improvement model because it measures the number of students a

school is able to move up from the lowest levels of achievement, even if they do not meet proficient levels (The Center for Public Education, 2007). The improvement model has been valued for being relatively easy to calculate and communicate to stakeholders; however, it does not reflect individual student growth or change within the same achievement levels (The Center for Public Education, 2007).

Growth Model

Whereas status and improvement models focus on measuring the achievement levels of groups of students, the growth model tracks "the achievement scores of the same students from one year to the next with the intent of determining whether or not, on average, the students made progress" (Goldschmidt et al., 2005, p. 4). One form of growth model is a simple growth model that details the difference in scale scores for a student between two years (The Center for Public Education, 2007). The simple growth model, however, does not measure whether a student is on track to become proficient in a given academic area; this is one reason why some states implement a growth to proficiency (O'Malley, Murphy, McLarty, Murphy, & McBride, 2011; The Center for Public Education, 2007) or growth to standard (Betenbenner, 2008) model that details not only the difference in achievement but also a student's trajectory toward achieving proficiency. Carey and Manwaring (2011) describe two growth models, trajectory and transition table, that appear to fit the description of a growth to proficiency model. Specifically, a trajectory model takes a student's current achievement level and requires that the gap to proficiency be closed over a three to four year period; a transition table monitors a student's growth in a below proficient level and does not necessarily require that students actually achieve proficiency as long as growth is made (Carey & Manwaring, 2011). Yet another type of growth

model is the student growth percentile that compares a student's growth with students who achieved similar scores in the past and then places the student in a percentile (0-100) to compare his/her growth with peers of similar academic backgrounds (Carey & Manwaring, 2011).

States and districts have come to value growth models such as these largely because they are more grounded in the belief that student learning is best reflected in changes in achievement as opposed to measuring whether a student meets a specific target or benchmark (Betenbenner & Linn, 2009; O'Malley et al., 2011). Criticisms of growth models focus mainly on how difficult it is for the general public to understand them and the threat to reliability should assessments change over time (Betenbenner & Linn, 2009; O'Malley et al., 2011).

Value-Added Model

The value-added model (VAM) is also a type of growth model but is unique in that it utilizes statistical methods to isolate the specific effects of a given teacher, program, school, or district on student achievement (Goldschmidt et al., 2005). VAM determines whether a student has made sufficient growth based on an estimate calculated using the student's past achievement scores (The Center for Public Education, 2007). Harris and McCaffery (2010) offer a well-respected definition of VAM, describing it as:

> … any analysis using longitudinal data to study the effects of educational inputs on achievement … [where] the potential student outcome is the average of the outcomes the student could have with other teachers who might teach the class. (pp. 253-254)

These techniques have also been described under the term "projection model" (Carey & Manwaring, 2011; O'Malley et al., 2011). Among the most praised aspects of VAM is the focus on student growth using criterion-referenced items (Peterson, 2010; Peterson & Peterson, 2006;

Sanders, Saxton, & Horn, 1997; Stiggins & Duke, 1988; Stronge & Tucker, 2000; Webster & Mendro, 1997) and the ability to control for factors outside of a teacher's control (Haertel, 1986; Harris & McCaffery, 2010; Thum & Bryk, 1997). Proponents of VAM argue that the sophistication of such analysis allow for a higher level of accountability for teachers and schools in terms of measuring their effects; however, there is also a large population of scholars who still question the accuracy and stability of VAM (Betenbenner, 2008; Betenbenner & Linn, 2009; Carey & Manwaring, 2011; Goldschmidt et al., 2005; O'Malley et al., 2011; Steele, Hamilton, & Stecher, 2010).

Virginia's Models for Measuring Student Learning

Virginia has adopted status, improvement, and student growth percentile models for the purposes of measuring and reporting student learning. Every school annually receives a state issued accreditation rating based on "overall achievement in English, history/social science, mathematics and science" (status model) and annual measurable objectives (AMOs) that measure the performance of key groups of students using targets based on those groups' previous performance on states assessments in reading and mathematics (VDOE, 2012). In addition, the Virginia Department of Education also provides school districts with student growth percentiles that compare an individual student's growth in reading and mathematics with the growth of other students who have similar past achievement levels on SOL tests (VDOE, 2012). While all school districts are required to use some measure of student academic progress as an integral part of teacher evaluation, they are not required to use student growth percentiles . Those schools participating in the Virginia Pay for Performance program are required to use student growth percentiles (VDOE, 2011c).

Introduction to *The Personnel Evaluation Standards*

One of the more helpful ways to examining how appropriate the use of student learning measures is in teacher evaluation is to do so through the lens of *The Personnel Evaluation Standards* (Gullickson, 2009). In response to a growing need to provide guidance for evaluation expectations in education, the Joint Committee on Standards for Educational Evaluation was formed in 1975 (Sanders, 1996). The committee ultimately created the first edition of *The Personnel Evaluation Standards* in 1988 (Gullickson, 2009), a document that has come to be adopted across the country as the central resource for developing, critiquing, and revising education evaluation policies and practices. The standards are organized into four main areas: Propriety, Utility, Feasibility, and Accuracy. Within each area is a series of sub-standards that provide policymakers and administrators a framework for examining evaluation practices.

The use of student learning measures in teacher evaluation is a relatively new one and it is also helpful to examine the literature on the topic through the lens of *The Personnel Evaluation Standards* (Gullickson, 2009). The existing literature has centered around three of the four areas: Propriety, Utility, and Accuracy. Since the Feasibility standards largely focused on whether or not a practice can be implemented, there is little written on the use of student learning measures in teacher evaluation presumably because it is such a recent trend. As such, this review of the literature will center on the three relevant clusters of standards.

Issues of Propriety

The Propriety Standards described in *The Personnel Evaluation Standards* (Gullickson, 2009) are meant to provide guidance so that an evaluation is legally defensible, ethically centered, and keenly focused on the welfare of the employee (Gullickson, 2009; Sanders, 1997).

Among the standards provided under the umbrella of propriety are: Service Orientation, Appropriate Policies and Procedures, Access to Evaluation Information, Interactions with Evaluatees, Comprehensive Evaluation, Conflict of Interest, and Legal Viability (Gullickson, 2009). Of the seven standards, three have a direct and relevant link to the use of student learning measures in the teacher evaluation process (see Table 1).

Service Orientation (Standard P1)

The service orientation standard states that the evaluation process should encourage the sound education of a school's students so that the needs of a school, community, and society are met (Sanders, 1996). Chief among the eight guidelines provided under the standard that relate to the use of achievement data are the following detailed by Gullickson (2009):

- Determine the purposes and uses of the evaluation that reflect the needs of the students and community and the roles and responsibilities of the evaluatee, then plan and conduct the evaluation to serve those needs.
- Ensure that evaluations serve to protect the rights of students for adequate instruction, service, and equal educational opportunity.

To analyze the use of achievement data in this context, it is helpful to look at how the practice either encourages or discourages sound educational practice as well as to examine the goal of equal academic opportunity for all students.

Sound education. Current reviews of relevant literature, policy briefings, and educational histories overwhelmingly express caution about the potential effects of using student learning measures in the teacher evaluation for the provision of a sound education for students (Braun et al., 2010; Corcoran, 2010; Darling-Hammond, 2009; Darling- Hammond & Rustique-

Forrester, 2005; Hinchey, 2010; Marshall, 2009; Misco, 2008; Mujis, 2005; Peterson, 2000; Ravitch, 2010).

Table 1

Propriety Standards Related to Use of Learning Measures in Evaluation

Standard	Description	Use in Framework	Reason
P1 Service Orientation	Personnel evaluations should promote sound education of all students, fulfillment of institutional missions, and effective performance of job responsibilities.	Yes	Student learning measures are a form of reporting on a sound education.
P2 Appropriate Policies and Procedures	Guidelines for personnel evaluations should be recorded and provided to the evaluatee in policy statements, negotiated agreements, or personnel evaluation manuals.	Yes	The use of student learning measures in teacher evaluation is relatively new and will require new policies and procedures.
P3 Access to Evaluation Information	Access to evaluation information should be limited to the people with established, legitimate permission to review and use the information.	No	Use of student growth percentiles presents no new challenges to confidentiality as they are provided through a secure web portal to administrators and it is a local decision on how it is shared.
P4 Interactions With Evaluatees	The evaluator should respect human dignity and act in a professional, considerate, and courteous manner.	No	Standard would be applied regardless of the use of student learning measures.
P5 Comprehensive Evaluation	Personnel evaluations should identify strengths and areas for growth.	Yes	Student learning measures may reflect new strengths or areas for growth not previously available to evaluators.
P6 Conflict of Interest	Existing and potential conflicts of interest should be identified and dealt with openly and honestly.	No	Standard would be applied regardless of the use of student learning measures.
P7 Legal Viability	Evaluations should meet the requirements of applicable laws, contracts, collective bargaining agreements, affirmative action policies, and local board or institutional policies.	No	The use of student growth profiles in teacher evaluation has previously been approved by state legislature. Virginia is also a right to work state.

Note: Adapted from *The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators.* (p. 27), by A. Gullickson, 2009, Thousand Oaks, CA: Corwin Press. Copyright 2009 by The Joint Committee on Standards for Educational Evaluation, Inc. Adapted with permission.

Chief among concerns is that student learning is measured with the use of mostly standardized tests that assess basic skills (Braun et al., 2010; Corcoran, 2010; Darling-Hammond, 2009; Darling-Hammond & Rustique-Forrester, 2005; Mujis, 2005; Peterson, 2000; Ravitch, 2010). Specific concerns regarding the use of standardized tests center on the potential to narrow the curriculum to focus on key factual knowledge as opposed to fostering critical thinking skills (Braun et al., 2010; Corcoran, 2010; Darling-Hammond, 2009; Darling-Hammond & Rustique-Forrester, 2005; Marshall, 2009; Misco, 2008; Mujis, 2005) as well as ignoring other subjects that may not be tested (Callier, 2010; Darling-Hammond, 2009; Darling-Hammond & Rustique-Forrester, 2005). In addition, such a focus on academic performance may also discourage teachers from addressing non-academic areas such as the teaching of social and behavioral skills that are often considered to be part of a sound education (Callier, 2010; Corcoran, 2010; Hinchey, 2010).

Much of the research supports the concerns articulated in the non-research based literature regarding how the use student learning measures in teacher evaluation can discourage a sound education. In one of the rare studies that was able to examine randomly-assigned teachers to classrooms in a large school district (156 classrooms, 3194 students), Kane and Staiger (2008) find that teacher effects on student achievement faded out by approximately 50% annually for the two years following a student's placement with a teacher, a finding that puts into question the long term effects of a teacher in the quest for a sound education. Further questioning about whether the practice of using learning measures in teacher evaluation promotes a sound education surfaces in a study of the Pennsylvania Value-Added Assessment (PVAAS). In examining the 93 school

districts implementing PVAAS, McCaffery & Hamilton (2007) conclude that PVAAS had no effect on student achievement. Despite the perceived lack of impact the PVAAS had on student achievement, 80% of the administrators surveyed believed that PVAAS measures what the district is doing to improve student achievement as opposed to 50% believing that AYP benchmarks set by NCLB do.

While much of the research paints a negative picture of the effects of using student learning measures in teacher evaluation as it relates to providing a sound education, these studies are often limited to a single school or school division. The Measures of Effective Teaching (MET) project sponsored by the Bill and Melinda Gates Foundation has articulated findings that may put many of the concerns stated in previous studies to rest. The project is remarkably large in scale, including 3000 teachers in six large school districts, and has partnered with 17 of the foremost researchers in the field to generate conclusions on how much "teachers matter" in providing a sound education to students. Chief among the findings is that teachers who score high on value-added assessments not only have students who have high pass rates on state standardized tests, but they also promote a deeper understanding of the material beyond factual recall (Cantrell & Kane, 2010). Teachers in general also have large effects on student achievement in both math and reading, but to a greater extent in math (Cantrell & Kane, 2010).

Equal academic opportunity. Much of the non-research based literature available on the topic also warns against the second aspect of the service orientation standard examined, the ability to adequately educate all children. One concern is that tying teacher evaluation to student achievement may discourage teachers from wanting to

teach needier students (Baker et al., 2009). Ravitch (2010) also considers the potential for "gaming the testing system" by taking part in practices such as refusing admittance of low-performing students, reducing the number of low-performing students who test, and allowing for more accommodations than what would be typically allowed. In the context of using student growth measures, as opposed to benchmark scores, teachers may focus heavily on those students who can show the most growth instead of providing the same amount of attention and instruction to all students (Peterson, 2000).

Research literature related to the topic, most of which examines the impact that a highly rated or ranked teacher has on the academic performance of key student subgroups, is mixed as it relates to the education of all students. One study of 250 secondary teachers and the rankings they earned through a value-added assessment found that rankings varied depending on the statistical model used, classes taught, and the year that the data was collected. In addition, variation in the rankings was largely correlated with student characteristics such as race and socio-economic status (Newton et al., 2010). Newton et al. (2010) argue that the finding may discourage teachers from teaching students who are more at risk. Evidence of such a phenomenon may exist in a study by Borman and Kimball (2005) that compares the achievement results of 7,000 students and standards-based evaluations scores of the 400 teachers who taught them in grades 4-6. Specifically, classrooms made up of students with higher percentages of minority, poor, and low-achieving students were more likely to be taught by a teacher who had received a low evaluation score. The study also analyzes the effect of "teacher quality" on closing the achievement gap and found no conclusive evidence to support the premise that higher quality teachers are more capable of closing these gaps. Konstantopoulos (2009)

examines the impact of an "effective teacher" as defined by the ability to create academic gains in a study of Project STAR in the late 1980's, and finds that some, but not all, groups of students benefit from these teachers. Wright, Horn, and Sanders (1997), however, conclude that effective teachers are effective with students of all performance levels regardless of classroom make-up in their seminal study of the Tennessee Value-Added System (TVAAS). More recently, a study with a large amount of longitudinal information on individual student scores from the UTD Texas School Project indicates that a high quality teacher can be particularly effective with students of low socio-economic status (Rivkin, Hanushek, & Kain, 2005). Similarly, Aaron, Barrow, and Sanders (2007) find the largest impact of a high quality teacher is on the achievement of African-American students in math.

Appropriate Policies and Procedures (Standard P2)

The Appropriate Policies and Procedures standard states that guidelines for evaluation should be available in policy statements, negotiated agreements, and evaluation manuals (Gullickson, 2009). Chief among the ten guidelines under the standard that relate to the use of achievement data are the following detailed by Gullickson (2009):

- Include written policies in all appropriate and accessible documents such as employee handbooks, memos of understanding, faculty agreements, and so forth.
- Ensure that policies address all the elements for effective personnel evaluation set forth in these standards and are aligned with the goals and mission of the organization.

- Clarify in writing the differences in performance expectations associated with identified personnel classification levels, so that evaluations are performed in accordance with the specific job expectations.
- Make guidelines sufficiently specific and clear to enhance common understanding.
- Change evaluation guidelines when evaluation practices are changed, when guidelines are in conflict with applicable law, or when role definitions change.

Literature related to whether policies and procedures centered on the use of student learning measures in teacher evaluation is extremely limited. Tucker and Desander (2006), in delineating the legal principles for teacher evaluation, emphasize that "criteria for evaluation and the exact procedures to be followed should be formalized in a written policy" (p. 93). In his analysis of some of the pitfalls related to the practice, Haertel (1986) advises that many of the potential problems with using student learning measures in the evaluation process can be overcome by detailing what appropriate procedures for data collection and analysis will be used, but he does not offer case studies or examples of this occurring. Similarly, in a handbook chapter detailing issues in the design of accountability systems, Linn (2005) calls for "well-defined" procedures for analyzing student achievement data.

Literature on the policies and procedures centered on using student learning outcomes in teacher evaluation appears to be an area of study that will be growing in the near future. In its annual review of state policies regarding teacher evaluation, the National Council for Teacher Quality (2011) states:

> Over this same short period of time [between 2009-2011], we've seen dramatic changes regarding the use of student achievement data to inform teacher evaluations. In 2009, 35 of the 50 states did not, even by the kindest of definitions, require teacher evaluations to include measures of student learning. Only four states could be said to be using student achievement as the preponderant criterion in how teacher performance was assessed. Today 23 states require that teacher evaluations include not just some attention to student learning, but objective evidence of student learning in the form of student growth and/or value-added data. (p. ii)

The same report concludes that since the movement to use student learning measures in evaluation is still in its infancy, it is too early to truly evaluate the policies related to it (NCTQ, 2011).

Virginia, being a state that has adopted policies after publication of the NCTQ (2011) report, has just recently provided its own language regarding the use of student learning measures in teacher evaluation. Specifically, Virginia law states:

> School boards shall develop a procedure for use by division superintendents and principals in evaluating instructional personnel that is appropriate to the tasks performed and addresses, among other things, student academic progress … (Code of Virginia, § 22.1-295)

In addition, Virginia has required that every school board:

> …shall provide a program of high-quality professional development (i) in the use and documentation of performance standards and evaluation criteria based on student academic progress and skills for teachers and administrators to clarify

roles and performance expectations and to facilitate the successful implementation of instructional programs that promote student achievement at the school and classroom levels. (Code of Virginia, § 22.1-253.13:5)

Current literature has not analyzed the individual school board policies of Virginia school districts and their alignment with state law.

Comprehensive Evaluation (Standard P5)

The Comprehensive Evaluation standard states that evaluations should address both strengths and weakness and be conducted in fair manner to the evaluatee (Gullickson, 2009). Chief among the six guidelines of the standard that are relevant to the use of achievement data in the teacher evaluation process are the following cited by Gullickson (2009):

- Ensure that the evaluation procedures allow for comprehensive and accurate indications of evaluatees' strengths and weaknesses.
- Ensure that evaluates know what will be assessed, how evaluation data will be collected, how evaluation information will be used to identify strengths and weaknesses, and how the evaluation results can be used to design appropriate follow-up actions.
- Describe and justify the basis for interpreting both positive and negative assessment information and results.
- Record incidents outside the evaluatees' control that might account for inadequate performance.

The following section of the review of relevant literature will focus on providing the evaluatee a record of strengths and weaknesses as that is a prevalent theme among the

guidelines. Other elements of this standard such as ensuring that the evaluatee is knowledgeable of the assessment practices falls under the umbrella of Appropriate Policies and Procedures. Similarly, issues regarding incidents beyond the evaluatees' control are more thoroughly and appropriately discussed in the Accuracy and Analysis of Context section of the literature review.

Little has been written on evaluators' direct use of student learning measures to identify a teacher's strengths and weaknesses. Teacher evaluators trained in specific observation protocols have shown the ability to identify a teacher's instructional strengths and weaknesses and use that information as a way to improve practice (Sartain, Stoelinga, & Brown, 2011; Stiggins & Duke, 1988). While student achievement data offers a reflection of one educational outcome, it may also provide too narrow of a glimpse of desired educational outcomes (Darling-Hammond, 2009). In a study of a Florida school district that compares the evaluation scores teachers received from their principals and their value-added scores, Harris and Sass (2009) note that the lack of correlation between the two may be largely because of other educational outcomes principals value in addition to student achievement. Examples of these other outcomes include lifelong learning and enjoyment of the educational experience (Stronge et al., 2007) as well as fostering the beliefs behind multicultural education (Borman & Kimball, 2005). In summary, student learning measures do provide a platform for identifying some instructional strengths and weaknesses, but the literature also warns that the scope and purposes of teaching should not be limited to just achievement outcomes.

Issues of Utility

The Utility Standards outlined in *The Personnel Evaluation Standards* are meant to ensure that an evaluation is sufficiently "informative, timely, and influential" (Gullickson, 2011, p. 69). The Utility Standards are comprised of six main standards including Constructive Orientation, Defined Uses, Evaluator Qualifications, Explicit Criteria, Functional Reporting, and Follow-up and Professional Development. The literature on the use of student achievement data in teacher evaluation addresses four of those standards (see Table 2).

Constructive Orientation (Standard U1)

The Constructive Orientation standard is meant to "help school districts develop their human resources and assist teachers to excel in the responsibilities" (Sanders, 1996, p. 96). The most recent edition of the standard (Gullickson, 2009) identifies eight key goals of the standard, four of which are addressed in the literature:

- Selection and retention of proficient personnel
- Reinforcement of good practice
- Promotion of professionalism
- Fostering of collegiality

These four areas will act as the framework for reviewing the literature as it related to Constructive Orientation.

Selection and retention of proficient personnel. The literature available on the use of achievement data in teacher evaluation as it relates to the selection and retention of

Table 2

Utility Standards Related to Use of Student Learning Measures in Evaluation

Standard	Description	Use in Framework	Reason
U1 Constructive Orientation	Personnel evaluation should help institutions develop human resources and encourage and assist evaluatees to perform in accordance with the institutions mission and goals.	Yes	Student learning measures could impact mission and goals and different measures could reflect different institutional priorities.
U2 Defined Uses	The intended users and uses should be identified in policies and guidelines at the beginning of an evaluation process.	No	Addressed in P2 Appropriate Policies and Procedures.
U3 Evaluator Qualifications	The evaluation system should be developed, implemented, and managed by people with the necessary skills, training, and authority.	Yes	Using student learning measures requires that evaluators develop a new skill set.
U4 Explicit Criteria	Systems of evaluation should have clear, specific criteria directly related to the required job expectations for the evaluatees.	Yes	Student learning measures provide the potential for a new set of criteria and expectations.
U5 Functional Reporting	Reports should be clear, timely, accurate, and germane to the purpose of the evaluation.	Yes	Student learning measures require reporting techniques not previously used in teacher evaluation.
U6 Follow-Up and Professional Development	Personnel evaluations should lead to appropriate professional development.	No	Addressed in P2 Appropriate Policies and Procedures and standard would largely be applied regardless of inclusion of student learning measures.

Note: Adapted from *The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators.* (p. 69), by A. Gullickson, 2009, Thousand Oaks, CA: Corwin Press. Copyright 2009 by The Joint Committee on Standards for Educational Evaluation, Inc. Adapted with permission.

teachers focuses on the staffing of high needs schools. Of particular note is the disincentive that using student performance as an evaluative tool in terms of building a proficient staff in a school comprised largely of high needs students, especially if student characteristics are not included in the measure of the value added by a teacher (Baker et al., 2010; Newton et al., 2010). This perception is reinforced by Kane and Staiger's (2008) observation that classrooms are rarely randomly assigned to teachers and, as a result teachers may want to avoid being assigned a class of students with lower previous achievement levels. In fact, the most proficient personnel are more often assigned to higher-achieving students as reflected in the correlations of achievement results and teacher standards-based evaluation scores by Borman and Kimball (2005). The implication of such a finding is that retaining a proficient teacher in a high needs environment is difficult in an evaluative setting that heavily weighs student achievement results. No literature was found indicating that use of student learning measures in an evaluative context aids in the selection and retention of personnel.

Reinforcement of good practice. When examining the reinforcement of good practice, it is important to differentiate between the use of student learning measures outside of the evaluative context and the use of them as a key component of the evaluation process. It has long been documented that a teacher's instructional practice can improve through the analysis and reflection on achievement data (Darling-Hammond & Rustique-Forrester, 2005; Stiggins, 2001). As the movement toward value-added assessment began to grow in the mid 1990's, leaders emphasized that the practice was one that could serve both summative and formative purposes (Sanders, Saxton, & Horn, 1997; Webster & Mendro, 1997). As the practice of using student achievement data

increased, scholars such as Darling-Hammond and Rustique-Forrester (2005) began to warn that such accountability measures have:

> …in some contexts, had positive influences on teaching and teacher quality; however, unintended negative consequences have also been found in systems that use limited measures and that emphasize sanctions without attention to improving school and teacher quality. (p. 290)

Ravitch (2010) concurs in stating:

> What was once an effort to improve the quality of education turned into an accounting strategy: Measure, then punish or reward … The strategy produced fear and obedience among educators; it often generated higher test scores. But it had nothing to do with education. (p. 16)

The warnings that Darling-Hammond (2005) and Ravitch (2010) offered are reinforced in other literature. For example, in one of the more influential pieces on teacher evaluation, Fernstermacher & Richardson (2005) emphasize that not only do instructional strategies need to be properly focused, but also ethically and morally acceptable. Teachers themselves have articulated that the focus of value-added assessment may not encourage these ideals but instead foster practices such as "teaching to the test" (Cooper, Ehrensal, & Bromme, 2005).

Promotion of professionalism. While making the use of achievement data central in the evaluation process may adversely affect instructional practice, there is also evidence that it can influence the professionalism of an employee and an institution. The focus on student achievement has led many to assume that schools will then revisit curricular alignment of courses and invest more heavily in professional development

centered around essential instructional practices (Darling-Hammond & Rustique-Forrester, 2005; Webster & Mendro, 1997). One such example comes from Vaughn Elementary School, a charter school established around an individualized teacher evaluation system that relies heavily on student achievement, where the entire teaching staff took part in 61 hours of professional development in literacy alone (Gallagher, 2004). The literature also warns of more negative influences of the use of achievement data in evaluation on the professionalism of teachers. Principals have been criticized for viewing teaching as a "semiprofession" previously and teachers say that this perception is perpetuated by their feeling of having to focus their instruction on mainly what a standardized test assesses directly (Cooper et al., 2005; Hanushek & Rivkin, 2010). In addition, the high stakes nature of making student achievement a central part of a teacher's evaluation have led some to be concerned about "gaming the testing system" (Ravitch, 2010) or even encouraging cheating on the tests (Hanushek & Rivkin, 2010).

Fostering of collegiality. The final goal of the constructive orientation standard affected by the use of student learning measures is fostering collegiality. A study of the relationship of teacher evaluation scores and student test results in a large Western school district suggests that one reason for the mixed results in finding a correlation is that principals may have been more focused on improving staff morale than using test data as a method of remediating or dismissing a teacher (Kimball, White, Milanowski, & Borman, 2004). This type of suggestion echoes concerns raised in Darling-Hammond, Wise, and Pease's (1983) relatively early review of teacher evaluation that emphasizes how outside accountability measures are in conflict with a school's desire to build trust. In reviews of different methods of using student achievement in the evaluation context,

both Schalock and Schalock (1997) and Cowart and Myton (1997) describe teachers as "hostile" toward the use of test scores. A more recent survey notes that teachers in general do not "trust test scores" (Rosenberg & Silva, 2012, p. 5). Similarly, a recent policy briefing warns that the practice can create a sense of competition among teachers and lead to a lack of collaboration (Baker et al., 2010).

Evaluator Qualifications (Standard U3)

The next relevant standard under the Utility area of *The Personnel Evaluation Standards* is that evaluators should been properly trained and skilled at evaluating a teacher (Gullickson, 2009). At the core of this standard is the understanding that a well-trained principal is the primary evaluator of teachers (Nolan & Hoover, 2004). As such, the evaluation process is often one where the "burden is carried alone" by the principal (Hallinger, 2003). This context of evaluation practices coupled with Stiggins' (2001) assertion that administrators are often not well versed in assessment literacy is cause for pause when considering the use of achievement data in teacher evaluation. A recent briefing paper also echoes the position that evaluation by competent evaluators must be at the center of teacher evaluation (Baker et al., 2010), but much of the research indicates that this is more of an aspiration than a statement of fact (Derrington, 2011; Stiggins, 2001). For example, an in-depth analysis of principal preparation programs across the country estimates that only about 2% of the course time that prospective school administrators took focused on accountability (Hess, 2007). Furthermore, training in the realm of teacher evaluation was much more focused on coaching and supervision than it was on "tough-minded" evaluation that includes the use of student performance data in evaluation (Hess, 2007). In addition, a recent brief by the National Governors

Association (2011) acknowledges that while teacher evaluation policies have changed to include aspects such as student learning outcomes, little attention has been paid to supporting and training principals so they can effectively evaluate teachers in these new contexts. A survey of a western school district highlighted some of the potential results of such a lack of training as the study indicated that principals were able to identify the top 10-20% of teachers but had difficulty identifying the middle range in terms of teacher effects on student achievement (Jacob & Lefrgen, 2006). Yet another study suggested, however, that principals' ability to identify teacher characteristics correlated to better student achievement is improving (Jacob & Walsh, 2010). Two other qualitative studies suggest that teachers have mixed perceptions regarding their principals' competence in the teacher evaluation process (Kimball, 2002; Zimmerman & Deckert-Pelton, 2003). An extreme case detailing administrators' lack of awareness of and facility with a value-added assessment system in Pennsylvania revealed that 28% of the principals studied did not even know that were taking part in a value-added system and 42% never saw a report (McCaffery & Hamilton, 2007). The recently released, final findings from the three-year MET project notes that administrators tend to rate their teachers higher than outside observers but also offers the recommendation that including multiple observers can improve reliability (Cantrell & Kane, 2013).

Explicit Criteria (Standard U4)

In addition to a constructive orientation, an evaluation practice should also be based in explicit criteria (Gullickson, 2009). In the context of using student learning measures in a teacher's evaluation, there is a need for all stakeholders to understand what the desired outcomes are (Peterson, 2000; Peterson & Peterson, 2006; Stiggins, 2001). A

survey of the Grade 2-6 classroom teachers in one school district made up of 13 elementary schools demonstrates how difficult it may be to agree upon achievement targets. Specifically, the correlation between principals' evaluation scores and student achievement data was not consistent largely due to the potential of principals focusing more on a pass rate than student growth (Jacob & Lefgren, 2006). A study of the Pennsylvania Value-Added Assessment System indicated that not only were the targets unclear, but less than 25% of the teachers even knew that their school was implementing the PVAAS (McCaffery & Hamilton, 2007). The lack of awareness of concrete targets in these examples echoes the suggestion of Kimball and Milanowski (2009) that evaluators are apt to use "gut-level feelings" more than hard data.

Functional Reporting (Standard U5)

Regardless of whether or not the criteria are explicit, a system for the functional reporting of the achievement data and its place in the evaluation system is also necessary (Gullickson, 2009). Efforts to streamline reporting systems have been known to go under constant revision, extending to over a decade's worth of work in at least one case (Felner, Bolton, Seitsinger, Brand, & Burns, 2008). Chief among the needs in reporting the role of student learning measures in the teacher evaluation process is the use of actual data as opposed to general descriptors of the data such as "satisfactory" or "needs improvement" (Peterson, 2000). Perhaps one of the largest impediments in doing this is time. For example, Stronge and Tucker (2000) recommend using a time frame that allows an evaluator to notice patterns in performance, something that looking at the achievement scores of a single year may not allow. Furthermore, many evaluation systems base

decisions on end of year tests when the results may not be readily available as the timing for summative evaluations usually does not align with the typical standardized testing schedule (Webster & Mendro, 1997).

Issues of Accuracy

By far, the most influential research on using student achievement data in teacher evaluation revolves around the Accuracy Standards in *The Personnel Evaluation Standards*. The Accuracy Standards are meant to ensure that an evaluation is "technically adequate and complete to produce sound information appropriate for the purpose of making sound judgments" (Gullickson, 2009, p. 115). The Accuracy Standards are comprised of eleven main standards, six of which are reflected in the literature being reviewed (see Table 3).

Valid Judgments (Standard A1)

The Valid Judgments standard centers on what may be the "single most important issue in personnel evaluation" (Gullickson, 2009, p. 117). Specifically, validity centers on one's ability to trust whether the judgment made about an employee's performance is a trustworthy one (Gullickson, 2009; Sanders, 1997). Considering the standard in the context of the use of student achievement data in teacher evaluation, it is helpful to consider what the research reveals about the correlation between an individual teacher's behavior and student achievement. As Linn (2005) reminds his audience in a handbook chapter, validity is a "matter of degree" and scholars have yet to agree on whether the degree of validity is sufficient or not (Ingavarson & Rowe, 2008).

Table 3

Accuracy Standards Related to Use of Student Learning Measures in Evaluation

Standard	Description	Use in Framework	Reason
A1 Valid Judgments	Personnel evaluations should promote valid judgments about the performance of the evaluatee that minimize risk of misinterpretation.	Yes	Considerable research on how student learning measures are or are not accurate reflections of quality teaching.
A2 Defined Expectations	The evaluatee's qualifications, role, and performance expectations should be defined clearly.	No	Included with U4 Explicit Criteria.
A3 Analysis of Context	Contextual variables that influence performance should be identified, described, and recorded.	Yes	Considerable research on influence of context on student learning outcomes.
A4 Documented Purposes / Procedures	The evaluation purposes and procedures, planned and actual, should be documented	No	Included with P2 Appropriate Policies and Procedures
A5 Defensible Information	The information collected for personnel evaluations should be defensible and aligned with evaluation criteria.	Yes	Student learning measures provide for a new type of information to be used.
A6 Reliable Information	Personnel Evaluation procedures should produce reliable information.	Yes	Considerable research on reliability of student learning measures.
A7 Systematic Data Control	The information collected, processed, and reported about evaluates should be reviewed systematically, corrected as appropriate, and kept secure.	No	Standard would be applied regardless of the use of student learning measures.
A8 Bias Identification and Management	The evaluation process should provide safeguards against bias.	No	Standard would be applied regardless of the use of student learning measures.
A9 Analysis of Information	The information collected for personnel evaluations should be analyzed systematically and accurately.	Yes	Student learning outcomes are new type of data requiring new analysis techniques
A10 Justified Conclusions	Conclusions about an evaluatee's performance should be justified explicitly to ensure that evaluates and others with a legitimate right to know can have confidence in them	No	Relevant issues included with U3 Evaluator Qualifications
A11 Metaevaluation	Personnel evaluation systems should be examined systematically at timely intervals using these and other appropriate standards to make necessary revisions.	No	Standard would be applied regardless of the use of student learning measures.

Note: Adapted from *The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators.* (p. 27), by A. Gullickson, 2009, Thousand Oaks, CA: Corwin Press. Copyright 2009 by The Joint Committee on Standards for Educational Evaluation, Inc. Adapted with permission.

Correlation of evaluation scores and achievement. One of the ways in which researchers have worked to establish a valid link between teacher behaviors and student outcomes is by examining the correlation between evaluation scores and observations with student achievement. In one such study of teachers in a charter school with an individualized evaluation system for teachers, Gallagher (2004) noted a "strong, positive, and statistically significant relationship between teacher evaluations scores and student achievement" (p. 105) with particular strength in literacy. Similarly, a study of science achievement at a middle school indicated students who were taught by teachers deemed "effective" by their evaluation scores earned higher achievement scores than those who were not "effective" (Johnson, Kahle, & Fargo, 2006). While these findings occur on the school level, more persuasive studies have occurred on the district level. Similar relationships between evaluation scores and achievement scores have been found in studies of the Chicago Public Schools where "a principal in a school with 50% of students performing at or above proficient levels gives ratings that are about 0.3 points (0.5 standard deviations) higher than in a school with only 20% proficient" (Jacob & Walsh, 2010, p. 447). A similar relationship between evaluation score and student achievement was found a "large Midwestern school district" with correlation scores of .27 for science, .32 for reading and .43 for mathematics (Milanowski, 2004); similarly, coefficients on mentor ratings and future student achievement in New York City were stronger in math (.015 and .016) than reading (Rockoff & Speroni, 2010). Such findings are supported by a study of a "large Western school district" where positive, statistically-significant relationships were found in almost half of the grade-test combinations studied (Kimball, White, Milanowski, & Borman, 2004). Whereas these studies examine student

achievement as opposed to growth, Kane, Taylor, and Wooten (2011) note that data from the Cincinnati Public Schools indicate that an increase of one evaluation score "is associated with one-seventh of a standard deviation increase in reading, and one-tenth of a standard deviation increase in math" (p. 58).

While research, in general, suggests a very strong correlation between teacher ratings and student test scores, there has been work that puts the link into question. For example, in a study of teachers in four North Carolina school districts who had achieved National Board Certification, an achievement many consider an indicator of high quality teaching, the assumption that they produce higher achievement gains was not supported (Stronge et al., 2007). In addition, while Gallagher (2004) noted a strong correlation in literacy achievement, the same could not be said for mathematics achievement, a finding that was similar to White's (2005) study of the Coventry Rhode Island school district. When the lens of student achievement focused on closing the achievement gap, as opposed to overall achievement, results were mixed in the effectiveness of "high" or "low" quality teachers as defined by their evaluation scores (Borman & Kimball, 2005).

Correlation between observable characteristics and achievement. Another general approach to identifying a valid link between student learning measures and teacher evaluation has been to examine correlations between student performance and observable teacher characteristics typically considered to be influential in creating student academic gains. The need to examine such a correlation has been emphasized by researchers such as Konstantopolous (2009) who, in using data from a 4-year, large-scale, randomized experiment focused on general "teacher effects," concluded that making generalizations about "teacher effects" was difficult without having access to these very

observable characteristics. While the commonly held belief that individual teachers make a substantial impact on student achievement has held true in many cases, the analysis of the characteristics of individual teachers does not provide overwhelming insight to a direct correlation between observable teacher characteristics such as years of experience and education level and student gains in achievement. For example, large-scale research based on longitudinal data of students in Texas has suggested that teachers have "powerful effects" on student achievement but little variation exists when breaking the data down by observable teacher characteristics (Rivkin, Hanushek, & Kain, 2005). Similar findings occur in a study of San Diego elementary schools (Koedel & Betts, 2005). When a correlation between teacher characteristics and achievement has been made, teacher experience has been the strongest indicator of student achievement (Rockoff, 2004).

Longitudinal data. In the discussion of linking teacher behaviors with student achievement, there has been a call to use more longitudinal data so that the effects of time can also be considered (Newton et al., 2010). Early studies of the Tennessee Value-Added Assessment System were largely responsible for engaging researchers more recently in the study of individual teachers' short-term and cumulative effects on students over time (Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997). Some of the decade's more influential research has centered on students' achievement in mathematics in a given year. Specifically, a one standard deviation increase in math teacher quality has raised math scores by 0.13 grade equivalents, a finding that has proven stable over time (Aaronson, Barrow, & Sander, 2005). Similarly, Rivkin et al. (2005) suggest that the impact of a math teacher one standard deviation higher has more impact than

reducing a class size by ten students. The ambitious MET project also has recently released compelling findings that by using the project's multiple measurement approach to identifying effective teachers in a given year, it is possible to predict which teachers will be effective in subsequent years. Specifically, the MET project created teacher effect estimates using its protocol and measured those against student achievement gains of classrooms that were randomly assigned the next year. The project concludes:

> … in both math and English language arts (ELA), the groups of teachers with greater predicted impacts on student achievement generally had greater actual impacts on student achievement following random assignment. Further, the actual impacts are approximately in line with the predicted impacts. We also found that teachers who we identified as being effective in promoting achievement on the state tests also generated larger gains on the supplemental tests administered in spring 2011. (Cantrell & Kane, 2013, p. 8)

Cross-culturally speaking, high quality teachers also have demonstrated a substantial effect on student achievement in the areas covered for the High School Entrance exam in Beijin, China (Lai, Sadoulet, & de Janvray, 2007). Recent research on the cumulative effects of teachers over time has been mixed. Konstantopoulos' (2009) data from Project STAR in the late 1980's suggests that a teacher's positive effect persists substantially on a year-to-year basis for some student groups but not all. A study of one middle school, however, found that "effective teaching" as defined by a specific observation protocol, increased achievement of all student groups (Johnson, Kahle, & Fargo, 2006). A much larger experimental design found that teacher effects dissipated by roughly 50% per year in the two years following a student's assignment to a teacher (Kane & Staiger, 2008). In

short, the last decade's research suggests there is evidence that teachers have a strong impact when examining annual measures of student achievement, but it is difficult to determine the on-going effects beyond a single year.

Analysis of Context (Standard A3)

As pivotal as understanding the issues regarding the validity in using student learning measures in evaluation, so is the analysis of context when engaging in the practice. The working assumption is that good teaching leads to good learning (Fenstermacher & Richardson, 2005). Many in the field have expressed concern, however, that caution must be taken in such an approach as there may not direct causality between the two and an analysis of context is necessary to understand other influencing factors (Fenstenmacher & Richardson, 2005; Haertel, 1986; Harris & McCaffery, 2010; Ingavarson & Rowe, 2008; Kennedy, 2010; McConney, Schalock, & Schalock, 1997; Peterson, 2000; Peterson, 2006; Stronge & Tucker, 2000). The literature addressing the need for an analysis on context centers mainly on the influence of the following on student outcomes: previous and concurrent teachers, school structure and schedule, school resources, student characteristics and demographics, and community or cultural influences.

Previous and concurrent teachers. Among the factors that influence student achievement outside of the control of the current teacher is the influence of both previous and concurrent instructors (Baker et al., 2010; Braun et al., 2010; Corcoran, 2010). In an argument for using teacher work samples in evaluation, Cowart and Myton (1997) go as far to say that it is "indefensible" to "connect learning gains to individual teachers" (p.16). This stance may be largely due to the beliefs that highly effective or ineffective

teachers contribute to students' achievement results years after they had the student in class (Baker et al., 2010; Braun et al., 2010; Darling-Hammond, 2009; Darling-Hammond & Rustique-Forrester, 2005). Interestingly, research is mixed on the extent of the impact previous teachers have on student achievement (Johnson et al., 2006; Kane & Steiger, 2008). Scholars also caution that a student's other teachers in a given year can influence academic growth (Callier, 2010; Darling-Hammond, 2009). For example, it is unclear how much of a student's achievement should be attributed to each teacher if he/she switched teachers mid-year or if he/she receives supplemental instruction in addition to the core instruction (Steele, Hamilton, & Stecher, 2010). A study of 18 schools with populations that have traditionally been labeled as "at-risk," ESOL, and mobile sheds a particularly bright light on the influence of concurrent teachers. Specifically, every school in the study reported that it engaged in some sort of "shared instruction" practice, meaning that more than one adult worked with students in a single subject area; in addition, all but three of the classroom teachers in all 18 schools reported using this type of approach (Valli, Croninger, & Walters, 2007). What the shared instruction looked like varied as almost 75% of the students received "supplemental instruction" in reading and 31% in mathematics. Almost two-thirds of the classrooms reported having more than one adult in the room providing instruction in both reading and math (Valli et al., 2007).

School scheduling. The literature base has also identified school scheduling, with an emphasis on student class assignment, as a key element to be considered in the potential correlation of test scores and teacher evaluation. The need to consider these factors is not new. For example, the Oregon Teacher Work Sample methodology of

teacher evaluation not only linked learning outcomes to teacher evaluation, but it also included descriptions of the classroom, school, and community to add context (Cowart & Myton, 1997).

The call for policy makers to carefully consider the impact of classroom diversity and size has been echoed more recently as well (Baker et al., 2010; Darling-Hammond, 2005). Early analysis of the Tennessee Value-Added System, however, indicated that class size and diversity were minor influences in developing a correlation between student gains and teacher assessment (Wright et al., 1997). More recently, researchers studying a teacher's impact on student achievement in high school math suggest that teacher ratings are stable over time despite fluctuation in class size and sorting (Aaronson et al., 2007). In her review of relevant literature on the topic, however, Darling-Hammond (2009) warns that the assignment of "students who may be exceptionally difficult to teach" and "whose scores on traditional tests are problematic to interpret" (p. 13) can easily skew the estimates for student gains with a given teacher. Darling-Hammond's (2009) warning largely stems from studies such as Ballou et al., (2004), Lockwood et al., (2007), McCaffery et al. (2003) and McCaffery, Sass, Lockwood, and Mihaly (2009) who demonstrate the sensitivity of student gain estimates especially as they relate to students in key sub-groups.

Analyses of scheduling have been challenging because of the non-random nature of scheduling. In one of the few studies that was able to capture data from randomly assigned classrooms, researchers concluded that past student achievement was a good indicator of future performance but analysis of classroom characteristics provided even more accurate predictions of student gains (Kane & Steiger, 2008). While the literature

focuses mostly on student assignment as an influential scheduling factor, there are certainly others the literature has not directly addressed as thoroughly such as amount of time devoted to specific subjects (Gallagher, 2004).

School resources. The varying degrees of school resources available to teachers and students is another area examined in the literature when considering the context and the use of student learning measures in teacher evaluation. A thorough meta-analysis of over sixty studies focused on the influence of school resources on achievement concludes that school resources and student achievement are not only linked, but linked strongly (Greenwald, Hedges, & Laine, 1996). Most recently, briefing papers and recommendations to policy makers have warned that even high quality teachers would struggle to meet their potential when resources are scarce (Baker et al., 2010; Darling-Hammond, 2009). Greenwald et al. (1996) indicate that per pupil expenditure has consistently had a strong relationship to achievement, particularly when the money is spent on creating smaller schools and smaller class sizes. Another way money could be spent, raising teacher pay, has not proven to be effective in raising achievement according to Hanuchek and Rivkin (2007), but they also warn that various methodological issues do not allow for generalization on the topic.

Student influence. When examining the contexts of the teaching-achievement relationship, one of the factors that the literature has also addressed is the role that the student plays; in fact, some argue that the amount of influence that a student has on achievement is a critical factor to be understood before committing to using achievement in a teacher's evaluation (Baker et al., 2010; Darling-Hammond, 2009; Fenstermacher & Richardson, 2005; Kane & Steiger, 2008). Frymier (1998) is among the first to comment

on the issue in the age of teacher accountability, arguing that by making teachers responsible for student behavior, schools are removing responsibility from the student, a factor that surely influences achievement. Mendro (1998) counters the argument in a response piece stating that schools are responsible for motivating students and creating an internal drive for success. Teachers and parents who participated in one recent survey believe that teachers and students were equally responsible for a student's learning, with no parents stating that the teacher should be solely responsible (Ballard & Bates, 2008). Complicating the issue is a study of 250 teachers and 3500 students focused on the students' judgments of a teacher's effectiveness. Interestingly, English teachers were rated more favorably when they had a greater ratio of females in class (Newton, Darling-Hammond, Haertel, & Thomas, 2010). Furthermore, teacher ratings were lower when the teachers taught traditionally at-risk population, even when these were controlled for in the statistical model; in addition, ratings increased when teachers had higher populations of Asian students or students with highly educated backgrounds (Newton et al., 2010). Despite the fact that early value-added assessment models like the DVAS (Dallas Value-Added System) adjusted scores for challenging populations (Thum & Byrk, 1997), many wonder if this is enough. For example, Ballou, Sanders, and Wright (2004) responded to criticism of a lack of explicit controls for student demographics by applying one on past data from the TVAAS, resulting in a negligible impact on estimates of teacher effects.

Community and culture. The final area examined by the literature in regards to the analysis of context when using student achievement data for teacher evaluation is the influence of community and culture (Fenstermacher & Richardson, 2005). Rivkin et al.

(2005) found in a study of "omitted or mismeasured variables" that there were "large enough differences in the quality of instruction in a way that rules out the possibility that the observed differences are driven by family factors" (p. 449). Despite statements such as this, Callier (2010) warns that students who have families that are able to afford supplemental services may allow for those students to succeed despite having an ineffective teacher. This concern is echoed in a recent briefing paper along with the additional warning that teachers who teach students from low-income families typically must address summer learning loss as well (Baker et al., 2010).

Defensible Information (Standard A5)

When considering that *The Personnel Evaluation Standards* stipulate that defensible information should be used in evaluation, examination of the tests used in a system that includes student performance data is necessary (Gullickson, 2009). Specifically, any such system should ensure that the assessments used are fair and valid (Stronge & Tucker, 2000), and the assessments should be adequately aligned with the curriculum (Stronge & Tucker, 2000; Webb, 2002). In addition there is added focus to measure not just benchmark performance, but student growth (Herman, Heritage, & Goldschmidt, 2011).

The most common type of assessment used and scrutinized is the standardized test, largely because of their ability to assess a wide range of information and allow for comparability of students and teachers (Mujis, 2005). There has been substantial concern that standardized tests are not suitable for use in the context of teacher evaluation, however, because they may not be aligned to local curricula and vary in terms of quality (Braun et al., 2002; Haertel, 1986; Mujis, 2005; Peterson, 2000; Peterson & Peterson,

2006; Ravitch, 2010). As a way to measure alignment, Webb (2002) created and utilized the Webb Alignment Process to determine that many state's standardized assessments have "issues" and that any analysis of alignment between standards and assessment is based largely on subjectivity. The murky nature of assessment measures also surfaces in a policy briefing based on data from New York City and Houston, where substantial variation in a teacher's value-added score occurred depending solely on which test was used to measure the impact of the teacher, particularly when measuring achievement in reading (Corcoran, 2010). Despite such concerns, a study that correlated every state's level of accountability and performance on the NAEP mathematics test found that when states emphasize the importance of performance on assessments such as standardized tests, achievement increases (Carnoy & Loeb, 2002).

Whereas standardized tests have been appreciated for their ability to assess student knowledge of a broad range of content, scaled tests are considered far more appropriate in the context of teacher evaluation according to the literature. As Haertel (1986) expresses, scaled tests more directly measure student growth in a specific area and application of knowledge as opposed to just factual recall. This focus on growth is a critical aspect of value-added modeling in teacher evaluation but must be approached with caution (Braun et al., 2010; Darling-Hammond, 2009; Herman, Heritage, & Goldschmidt, 2011; Peterson & Peterson, 2006). Darling-Hammond (2009) has been among the most vocal in stating specific concerns citing the fact that most states have not developed such tests and the nature of content specific courses such as science or social studies, especially in secondary schools, does not encourage value-added approach as much as skills based subject areas like reading and mathematics. Even in a skills-based

subjects such as the performing arts or physical education, any sort of common assessment that measures growth from year to year is rarely available (Darling-Hammond, 2009). Even when a scaled test is available, a recent study of one district's value-added estimates of teachers in reading fluctuated depending on which type of scaled test was used (Papay, 2011).

Scholars have considered other assessment types as well but have quickly dismissed them for obvious reasons. For example, while Haertel (1986) indicated early in the discussion of using achievement data in teacher evaluation that norm-referenced tests were the only option to be used because of availability, others such as Stiggins and Duke (1988) were quick to point out how norm-referenced tests are too imprecise and the results are influenced by too many outside factors to be considered in such a high stakes environment. Peterson (2000) considers the use of locally constructed assessments but also emphasizes that extensive validity and reliability testing would be required, something that may not be a realistic possibility in most cases.

Reliable Information (Standard A6)

In addition to being defensible, the information used in a teacher evaluation must also be reliable according to *The Personnel Evaluation Standards* (Gullickson, 2009). The literature examines reliability from two angles. The first angle examines whether the information used is reliable from teacher to teacher, regardless of teaching assignment or area of specialty. The second considers whether the information is reliable on a year-to-year basis.

Availability of data. At the heart of considering whether the information for teacher evaluation is reliable from teacher to teacher is the very availability of student

achievement data. The estimates of the percentage of teachers who have appropriate student achievement data to be used varies from 60-70% (Webster & Mendro, 1997) 50% (Peterson, 2000) to 30% of elementary teachers and merely 10% of high school teachers (Darling-Hammond, 2009). Policy briefings and reviews of the practice have expressed how teachers of many subject areas including some sciences, social studies, fine arts, vocational education, and physical education do not have state tests (Corcoran, 2010; Darling-Hammond, 2009; Harris & Sass, 2009). Similarly, teachers of different grade levels are not covered equally by state tests (Darling-Hammond, 2009). Complicating matters is Callier's (2010) observation that elementary teachers typically teach all major content areas, which means that should a teacher score poorly in one area but not the others, it puts into question whether the teacher is effective or not. In addition, specialty teachers such as ELL and special education teachers typically do not have statewide student assessments suitable for the teacher evaluation context (Holdeheide, Goe, Croft, & Reschly, 2010).

Stability over time. The literature also points out the importance of stability over time when using student achievement data in teacher evaluation. A recent study of 250 secondary teachers and 3500 students indicated that the judgments of teacher effectiveness were largely dependent on which year's data were analyzed (Newton et al. , 2010). Similar instability was found in Corcoran's (2010) study of New York City's Teacher Data Initiative and Houston's ASPIRE program. Specifically, while many teachers who performed in the bottom quintile one year performed in the same quintile the next, as many as 23% of the teachers in the lowest quintile performed in the top two quintiles the next and 23% of the highest performers in one year performed in the bottom

two quintiles the next (Corcoran, 2010). Such fluctuations have suggested that a minimum of three years' worth of data should be used in teacher evaluation to account for variability (Darling-Hammond, 2009). Variability in teacher effectiveness ratings does not occur in all studies, however, as evidenced in Aaronson et al.'s (2007) assertion that the teacher ratings created in his study of Chicago public school's 9th grade math teachers maintained stability over time. Specifically, teachers in the lowest quality quartile were most likely to stay in that quartile the next year (36%) with 29% moving into the next quality quartile and 26% into the third quality quartile; conversely, teachers in the highest quarter were most likely to stay there the next year (57%) (Aaronson et al., 2007).

Analysis of Information (Standard A9)

When examining how the literature has framed the accuracy standards in terms of using student performance in teacher evaluation, the final relevant area is the analysis of information (Gullickson, 2009). In today's context of teacher evaluation, this is particularly important as it requires scholars and practitioners to critique the statistical models used in value-added assessment. In addition, the literature has identified the handling of missing data in a value-added model as a critical part of maintaining accuracy in judgments.

Value-added. Early discussion of the use of student achievement in teacher evaluation focused on standardized test performance (Haertel, 1986); however most now consider a value-added approach to be more appropriate (Peterson & Peterson, 2006; Hanushek & Rivkin, 2010). The movement toward value-added assessment, while often appreciated in theory and for its use with large data sets, has generated substantial debate

about the stability of the model when focused on individual teacher effects (Amrein-Beardsley, 2008; Baker et al., 2010; Braun et al., 2010; Darling-Hammond, 2009). For example, a study of the statistical model used to rank-order teachers in San Diego elementary schools suggests that the model is unstable due to a "relatively low signal-to-noise ratio" (Koedel & Betts, 2005). Similarly, a multilevel linear crossed random effects growth model has been criticized for the bias it is subject to when estimating teacher effects (Palardy, 2010). Interestingly, analysis of the variance created in different value-added models as opposed to the variance created by different mathematics assessments has indicated that the different models provide more stability than the different assessments (Lockwood, McCafferey, Hamilton, Stecher, Le, & Martinez, 2007).

Controls. When considering the stability of statistical models, the literature identifies the use or non-use of controls as a core issue. Darling-Hammond (2009) has asserted that value-added modeling is not sustainable because of the different pictures that the models paint of teacher effects depending on the controls used for students that are traditionally more difficult to teach (e.g., ESOL, truant, or homeless children). Such claims have been buoyed by studies such as Newton et al. (2010) that examines the data of a large sample of teachers and how the variance of teacher effectiveness is sensitive to whether or not student demographics and school fixed effects are controlled in the modeling. Lockwood et al. (2007) echo these findings in their analysis of longitudinal data from a large school district that indicates that estimates of teacher effects are highly influenced by the degree of controls applied in the calculations. The literature is not in total agreement, however, about the explicit use of controls for student demographics as noted by the work of Ballou et al. (2004) who in response to criticism of the TVAAS's

lack of controls for socioeconomic and other background factors applied controls for such; ultimately, they suggest that the controls had a "negligible" impact on estimating teacher effects.

In addition to the inclusion or exclusion of controls for student characteristics, the method of handling missing data has also become a source of debate. Similar to her assertions that controls act as a determining factor in creating teacher effect estimates, Darling-Hammond (2009) has made the same claim regarding missing data. Two recent reports meant for policymakers have advised them to simply not include students who do not have prior test scores in the calculations of estimated teacher effects, despite the potential for these students accounting for a large number in the overall population (Corcoran, 2010; Steele et al., 2010). Such an approach has not been advocated by all. An early description of the TVAAS notes that the system "enables a repeated-measures, multivariate response analysis allowing the inclusion of all of the information available for each student regardless of the degree of missing information" (Sanders et al., 1997, p. 137).

Summary of the Literature Review

A review of the literature provided some insight into the benefits and difficulties of using student achievement data in the context of teacher evaluation. Using *The Personnel Evaluation Standards* as a framework, three of the four domains (Propriety, Utility, and Accuracy) provide an umbrella for all of the standards that have relevance to the practice. For each, the literature has proven to provide little consensus on the appropriateness of using student performance on standardized tests in evaluation.

In terms of propriety, the literature has indicated that there are both some extremely valuable and potentially detrimental aspects of using achievement data in teacher evaluation. Specifically, the practice appears to have the ability to promote aligned curricula and assessments, focus attention on students with the most need, and provide a core focus on achievement as the purpose of teaching. There is concern, however, especially when one considers the number of circumstances outside of a teacher's control that can affect achievement. In addition, having such a strong focus on achievement has led many to ignore the other "jobs" that a teacher often embraces: counselor, mentor, and advocate.

Regarding utility, the literature has focused largely on the "constructive orientation" of using student achievement in teacher evaluation. While there is evidence of the practice improving teaching in some cases, there are also serious concerns about how teachers may begin to "work the system" and that collegiality may be an unintentional victim in an effort to encourage accountability. In addition, there is a wide range of variation in what the achievement expectations of a student should be, leaving teachers unsure of what the actual achievement targets are and what is expected of them.

Without a doubt, the area that has caused the greatest debate is whether student performance is an accurate portrayal of the work that teachers do. While most would agree that good teaching leads to learning, the debate emerges when policy makers assume that learning leads to achievement on a specific assessment. Furthermore, how to handle the influences of such factors as previous or concurrent teachers, school effects, community and culture, and availability of resources has complicated the issue. Finally,

there is no consensus on what is the most appropriate way to analyze student data once it is gathered.

Despite the number of concerns voiced in the literature, federal and state policies requiring the use student learning measures in teacher evaluation have increased. As a result, districts, schools, and teachers have been placed in a position where they are to make sense of an extremely complicated practice. There has been little research completed on teachers' perceptions of the appropriateness of using student learning measures in their evaluation. This study was designed to provide exactly that, in an effort to inform the process and the professional development that educators in a variety of contexts may need to provide as the movement gains more momentum.

CHAPTER 3: METHODOLOGY

Overview

A review of the literature on the use of student achievement data in teacher evaluation revealed that there are gaps in our knowledge of the practice. Specifically, there has been little research conducted on teachers' perceptions of the practice. In addition, the topic has not been examined through the lens of *The Personnel Evaluation Standards* (Gullickson, 2009), considered by most to be the authoritative set of standards for creating and reviewing education personnel evaluation practices. This study was designed to capture insight in regards to the practice of using student achievement data in teacher evaluation considering these two critical perspectives. By gathering data from teachers, this study complements the research literature and provides school administrators important information that can guide their evaluation practices as well as the creation of professional development opportunities. In addition, by using *The Personnel Evaluation Standards* as the framework (Gullickson, 2009), the study provides insight to the perceived levels of propriety, utility, and accuracy of the practice. The central research questions that guide the study are:

1. What are teachers' perceptions of the use of student learning measures in teacher evaluation in terms of propriety?

2. What are teachers' perceptions of the use of student learning measures in terms of utility?
3. What are teachers' perceptions of the use of student learning measures in terms of accuracy?

This chapter details the methods used in the study, specifically describing the context of the study, participants, instrumentation, data collection, data analysis, and limitations.

Context

The Commonwealth of Virginia is one of the many states employing the use of student achievement data in teacher evaluation. As a review of the literature indicates, the use of student learning measures is becoming one of the more popular and controversial methods for tackling the challenge of including student performance in this context. Beginning in 2011, the Virginia Department of Education began providing school districts student growth percentiles. The student growth percentiles are available for all teachers of students in grades 4-9 in reading and math. While the VDOE provides Student growth percentiles to all school districts in the Commonwealth, it is unclear how many actually use them in the teacher evaluation process. What is known, however, is that all schools taking part in a separate opt-in initiative, the Virginia Pay for Performance program, are required to use Student growth percentiles in teacher evaluation as part of the agreement.

Participants

The participants in this study came from two populations. The first population was the group of teachers who taught reading and/or math in grades 4-8 and whose schools took part in the Virginia Pay for Performance program in 2011-2012, and who

continued to be employed by the same school or school division in 2012-2013. Participants were identified by information available on the individual schools' web sites. This population was been identified because it is the only known group of teachers who received a summative evaluation that included student growth percentiles. In addition, while some 9th grade math teachers were evaluated using student growth percentiles, school web sites did not clearly delineate which math teachers were 9th grade math teachers and which were not. The second population was a collection of teachers in like schools who may or may not have received summative evaluations using student growth percentiles and who may or may not have been eligible to receive additional compensation based on their students' academic performance. This population was been identified because it provided an opportunity to compare the perceptions of teachers who are known to have received summative evaluations using student growth percentiles and/or who were eligible to receive additional pay, and teachers who may not have.

The first population was defined using the following information. In 2011-2012, 25 schools in 13 school districts were selected to participate in the Virginia Performance-Pay Pilot. Of those 25 schools, 9 were high schools (grades 9-12) and 1 was a K-2 school, leaving 15 schools that employed teachers of math and reading in grades 4-8. One middle school did not identify which teachers were math or reading teachers. The result is a population of 14 schools that publically shared which teachers taught math or reading in Grades 4-8 in 2011-2012, a total of 154 teachers. In 2012-2013, 110 of the 154 teachers returned to their same school. An additional 21 teachers were part of a division restructuring that required they move schools but stayed employed by the division. This resulted in a final population of 131 teachers that are publically known to

have taught reading or math in Grades 4-8 for a school that has taken part in the Virginia Performance-Pay Pilot, and who have received a summative evaluation using student growth percentiles. With such a small population, it was necessary to survey all 131 teachers.

The second population was determined by identifying schools that are similar to first population. When available, a school in the same division was chosen; when not available, a school in a neighboring district was used. In order to further identify like schools, community type (urban, suburban, rural), school size, and achievement as identified by state accreditation and AMO status was used. The second comparison population was close to double in size of the first with 268 teachers.

Approval

Approval of University of Virginia Institutional Review Board for the Social and Behavioral Sciences (see Appendix C) was obtained on April 23, 2013. In addition, all participants were presented with an online informed consent agreement at the beginning of the survey.

Instrumentation

An on-line survey (see Appendix B) was created to gather teacher perceptions on the use of student growth percentiles in teacher evaluation. The first part of the survey identified key demographic information that informed how the use of student growth percentiles was understood by teachers of different backgrounds. Specifically, the survey gathered information on grade level taught, subject area(s) taught, years of experience, professional organization affiliation, previous evaluation experience, participation in the Virginia Pay for Performance program, and eligibility for additional pay. The second

part of the survey asked participants to rate the extent to which they agree or disagree with statements pertaining to the use of student growth percentiles in teacher evaluation. The statements aligned with the relevant standards and guidelines set in *The Personnel Evaluation Standards* with a focus on propriety, utility, and accuracy (Gullickson, 2009). For each of the three clusters of standards, there was an open-ended section for participants to articulate further beliefs and opinions that were not captured by the survey questions.

Specific attention was paid to developing a survey that was simple in design. Surveys considered to be plain have been shown to receive a higher participation rate than colored ones (Dillman, Tortora, Conradt, & Bowker, 1998). A more recent study of on-line surveys revealed an increase of five percentage points in participation for simple surveys over their complex counterparts (Whitcomb & Porter, 2004). Considering the relatively small population to be surveyed, it was particularly important to value the impact simplicity can have on response rate.

In order to limit the amount of measurement error introduced into the study, the survey was evaluated in three ways prior to implementation. The first evaluation was through expert review with a focus on the following elements emphasized by Groves, Fowler, Couper, Lepkowski, Singer, & Tournangeau (2009): wording of the questions, structure of the questions, response alternatives, order of the questions, and instructions to the participants. Expert reviewers included current teachers and researchers in teacher evaluation. Results of the expert reviews were used to revise aspects of the survey prior to administration. Specifically, a description of the purpose of the survey and the definition of student growth percentiles was added to each section to assist participants in

understanding the specific intent of the survey items. In addition, some questions were reworded in order to eliminate unnecessary jargon. The second evaluation was through the use of cognitive interviews to understand how respondents understand the questions and create responses (Borg & Gall, 1989) using procedures outlined in Groves et al. (2009). Participants in the cognitive interview also included current teachers and researchers in the field. Finally, a pilot was administered as recommended by Fogelman & Comber (2007). Results of both the cognitive interviews and pilot were used in revising the instrument and process for administration. Specifically, there were some minor wording changes to statements on the survey and glitches involving the uploading of email addresses to the survey program were eliminated.

Data Collection

The survey was administered as an on-line survey with a link to the survey sent via e-mail following Dillman's Total Design Survey Method (1999). Much has been written about the potential pitfalls of on-line surveys as they relate to response rate with a focus on particpants' access to the internet being a chief concern (Couper, 2000; Dillman, Tortura, & Bowker, 1999; Evans & Mathur, 2005). Such concerns should be alleviated with not only the general knowledge that internet access has increased since the writings but also that the participants were identified through information provided by their schools' respective web sites, indicating that internet access was readily available to participants.

Chief among the reasons for the use of a survey was that it provides for anonymity, a particular concern when surveying teachers on such a sensitive topic. Pryor (2004) states that anonymity in surveys allows for a realistic reflection of participants'

perceptions. Considering that the topic of the study centers on methods used for participants' job performance evaluation and the pilot status of the Pay for Performance Program, a focus on anonymity was warranted to protect teachers from retribution for possible criticism. In addition, surveys provide for timely data collection and a low degree of researcher bias (Babbie, 2001; Krathwol, 1998), two important aspects considering the recent implementation of student growth percentiles and the potential that the practice holds in the near future. Particular attention was paid to attaining a reasonable and respectable response rate. Porter (2004) identifies multiple contacts, shorter survey length, incentives, and salience as key elements in increasing response rate. In light of these recommendations, each participant was contacted using e-mail and paper messages over a six week period. Using multiple methods of contacting participants has been shown to increase response rates (Kapolitz, Hadlock, & Levine, 2004) as has follow-up mailings (Larson & Chow, 2003). In addition, the survey was designed to take approximately 10-15 minutes to complete. In regards to salience, Porter (2004) states, "Salience is simply how important or relevant a survey topic is to the survey recipient" (p. 14), a characteristic that the survey topic likely had for the participants.

Data Analysis

To answer each research question, a composite score for the domains of propriety, utility, and accuracy (Gullickson, 2009) was calculated. Descriptive statistics including mean and standard deviation for each composite score were calculated. Mean and standard deviation for each question under a domain were also reported. In order to determine the predictive value of key independent variables such as years of experience,

affiliation with a professional organization, previous experience being evaluated with student growth percentiles, eligibility to receive additional compensation, and participation in the Virginia Pay for Performance Program, a regression analysis was run on composite scores. Finally, open-ended responses were analyzed through content analysis based on themes that emerge from the responses as recommended by Fink (2002), allowing for further understanding of teachers' perceptions as they relate to each domain and research question. As themes emerged, comments were categorized by the most significant domain (propriety, utility, or accuracy) studied. Once the closed and open ended responses were analyzed, the information found in them were used to answer the research questions.

Validity

Groves et al. (2009) state that validity "is the extent to which the survey measure accurately reflects the intended construct" (p. 274). Efforts to ensure validity in this study came in three forms. First, the survey was constructed after a thorough review of the literature related to the topic. Second, individual items on the survey were reviewed and checked to ensure that they measure the intended areas for the study. The core areas of study were taken from *The Personnel Evaluation Standards* (Gullickson, 2009), considered to be the most respected guide to reviewing and evaluating personnel procedures and policies. Finally, the survey went through a series of three evaluations: expert review, cognitive interviews, and a pilot.

Limitations

A major limitation of the study was that it focused on a small population of teachers. The population used for the study included every returning teacher in Grades 4-

8 who received a summative evaluation in 2011-2012 using student growth percentiles as required by their school's participation in the Performance for Pay initiative set by the Virginia Department of Education, and teachers in like schools who may or may not have been evaluated using student growth percentiles and/or who may or may not have been eligible for additional compensation based on student performance. The population was a rich one as half of the participants were a unique and select group of teachers; however, considering the small population it would be difficult to make generalizations regarding the use of student learning measures in teacher evaluation in other states or in schools that used different measures of student learning gains other than student growth percentiles.

Summary

This study aimed to understand the perceptions of teachers who received a summative evaluation by an evaluator who used student growth percentiles as a method for measuring student achievement under the teacher's tutelage. In 2011-2012, 154 teachers in Grades 4-8 were known to have been evaluated in this context because of their school's participation in a Performance for Pay initiative set by the Virginia Department of Education. Of the 154 teachers, 131 returned to the same or similar position and are available for participation in this study. To compare the perceptions of teachers who were part of the Virginia Pay for Performance program and those who may not have been evaluated using student growth percentiles or who may not have been eligible for additional pay, a second sample of 268 teachers was determined. An on-line survey framed around themes and standards set by *The Personnel Evaluation Standards* (Gullickson, 2009) were used to assess these teacher's beliefs and attitudes regarding the

use of student growth percentiles in teacher evaluation. The results of the survey shed light on one of the specific approaches of using student achievement data in teacher evaluation promoted by the Virginia Department of Education, and the use of student learning measures in general as it pertains to teacher evaluation.

CHAPTER 4

FINDINGS OF THE STUDY

Overview

This chapter begins with a description of the survey administration and demographic information of the participants. It then details the findings of the study as they relate to the three research questions:

1. What are teachers' perceptions of the use of student growth percentiles in teacher evaluation in terms of propriety?
2. What are teachers' perceptions of the use of student growth percentiles in terms of utility?
3. What are teachers' perceptions of the use of student growth percentiles in terms of accuracy?

The research questions were answered through an analysis of survey data that included both closed and open-ended responses. For closed-ended questions, composite scores were calculated based on the three main domains reflected in the research questions (Gullickson, 2009). A regression analysis was then run in order to determine the predictive value of key independent variables. In addition, means and standard deviations of individual items were calculated. For open-ended questions, the responses were coded and categorized around central themes as recommended by Fink (2002). The

themes were then assigned to one of the three domains reflected in the research questions (Gullickson, 2009) with the comments analyzed and described by question.

Participants

Participants in the study were selected from two groups. The first is the group of teachers who taught reading and/or math in grades 4-8 and whose schools took part in the Virginia Pay for Performance program in 2011-2012, and who continued to be employed by the same school or school division in 2012-2013. The second was a collection of teachers in similar schools based on district demographics who may or may not have received summative evaluations using student growth percentiles. The first group was comprised of 131 teachers and the second was comprised 268 teachers, resulting in a total sampling frame of 399 teachers.

All 399 members of the sampling frame received a personalized pre-notification letter informing them of the their participation in the study. Within approximately 3-5 business days of their receipt of the pre-notification letter, participants received an email with a direct link to the survey. Follow-up emails were delivered weekly for four weeks in order to encourage a larger return rate. One hundred eighty three participants began the survey with 150 ultimately completing it, resulting in a 46 per cent response rate and 38 per cent completion rate. Since the study specifically examines three research questions where the results will be compared, only completed surveys were analyzed.

Participant demographics reflect grade-level diversity. Table 4 details the grade levels taught by participants.

Table 4

Grade Level Taught by Participants in 2011-2012

4th Grade	24%
5th Grade	19%
6th Grade	20%
7th Grade	18%
8th Grade	17%
Did not teach in 2011-2012	1%

There was also diversity reflected in the subject area taught by participants as detailed in Table 5.

Table 5

Subject Area Taught by Participants in 2011-2012

Reading	34%
Math	48%
Reading and Math	16%
Did not teach in 2011-2012	2%

Participants also represented varying degrees of teaching experience as reflected in their reported number of years teaching in Table 6.

Table 6

Reported Number of Years Teaching by Participants

Less than 3 years	8%
3-5 years	13%
6-10 years	25%
11-20 years	29%
20 or more year	25%

Of the 150 participants who completed the survey, 44 per cent also reported being a part of a professional organization such as the NEA or AFT.

This study focuses particularly on the use of student growth percentiles in teacher evaluation. Thirty-six per cent of the participants reported that their evaluator used student growth percentiles in their evaluation the previous year; of the remaining participants, 41% reported that their evaluator did not use them and 21% reported not knowing if student growth percentiles were used in their evaluation. In addition, 13% reported that they were eligible for extra pay based on their summative evaluation, but another 13% also reported that they did not know if they were eligible or not. In regards to the Virginia Pay for Performance program, 11% of participants reported that their school was a part of it but 34% reported not knowing. This is a particularly noteworthy statistic as one-third of the sample was selected because they were from schools that were publicly reported as being part of the Pay for Performance program.

Research Question 1: Propriety

The first research question asks: What are teachers' perceptions of the use of student growth percentiles in teacher evaluation in terms of propriety? To address this question, participants rated their level of agreement on a Likert scale (1 = *Strong Disagreement*, 2= *Disagreement*, 3 = *Agreement*, 4= *Strong Agreement*) regarding statements about the use of student growth percentiles as it relates to propriety. Three statements were categorized under the category of propriety.

For the three statements, a composite score was calculated. As noted in Table 29 (see Appendix), the value of Cronbach's alpha was .878. With a potential minimum of 3 and a potential maximum of 12, a mean of 7.5 would represent an overall balance of agreement and disagreement. The mean composite score for propriety was 6.5 which indicates an overall perception of disagreement on the use of student growth percentiles

in teacher evaluation as it relates to propriety. Participants' responses by question are reported in Table 7.

Table 7

Teachers' Perception of the Use of Student Growth Percentiles as it Relates to Propriety

Factor	N	Min.	Max.	Mean	Std. Dev.
Promotes sound education	150	1	4	2.30	.915
Promotes education of all students	150	1	4	2.21	.963
Reflects teacher's strengths and weaknesses	148	1	4	1.97	.903

Assuming a mean of 2.5 would represent a balance of agreement and disagreement, it is noteworthy that all three items resulted in a mean lower than 2.5, indicating a general negative attitude toward the practice as it relates to each item. The lowest mean (1.97) was associated with identifying a teacher's strengths and weaknesses. As indicated in Table 20 and Table 22 in Appendix A, this item was also the 6th lowest mean of the 21 individual items analyzed in the study. As noted in Table 21 and Table 22 in Appendix A, the item of promoting a sounds education produced the 6th highest mean among all statements but it is still below what would be considered a positive rating.

To determine the predictive value of key independent variables, a regression analysis was run on the propriety composite scores. To do so, independent variables were coded into two categories for each independent variable studied. Table 8 summarizes how the variables were coded along with the reasoning.

Table 8

Categorical Coding for Regression Analysis

Ind. Variable	Category 1 (Coded as 0)	Category 2 (Coded as 1)	Reason
Experience	< 3 years 3-5 years	6-10 years 10-20 years > 20 years	Teachers in years 0-5 are more likely to be on an annual contract and often have a greater comfort level with data-based decision making.
Professional Affiliation	Yes	No	Already limited to two categories.
Previous Evaluation Experience	Yes	No Don't Know Didn't Teach	Perceptions of participants in "No", "Don't Know" and "Didn't Teach in 2011-2012" categories were not influenced by previous evaluation experience.
Eligibility to Receive Additional Pay	Yes	No Don't Know Didn't Teach	Perceptions of participants in "No", "Don't Know" and "Didn't Teach in 2011-2012" categories were not influenced by potential for extra pay.
Participation in Pay for Performance Program	Yes	No Don't Know Didn't Teach	Perceptions of participants in "No", "Don't Know" and "Didn't Teach in 2011-2012" categories were not influenced by participation.

As shown in Table 9, the R Square value of .100 in the model summary of the regression analysis indicates that 10% of the variance in composite score means can be explained using the model.

Table 9

Model Summary of Propriety Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.316	.100	.069	2.424

In addition, Table 10 confirms the model to be a good fit with a significance of .009.

Table 10

ANOVA of Propriety Regression Model

Model	Sum of Squares	df	Mean Square	F	Sig.
1	93.323	5	18.665	3.179	.009
	839.926	143	5.874		
	933.248	148			

Considering the significance level, it is then worth examining the individual coefficients of the regression as detailed in Table 11.

Table 11

Regression Table for Propriety Composite Scores

	B	SE B	β	t	Sig.
(Constant)	10.972	1.826		6.009	.000
Experience	-.569	.302	-.153	-1.880	.062
Professional Affiliation	.029	.408	.006	.070	.944
Previous Evaluation Experience	-.335	.452	-.064	-.741	.460
Eligibility for Additional Pay	-1.888	.638	-.252	-2.959	.004
Participation in Pay for Performance Program	.432	.638	.055	.677	.499

Of the coefficients, the only statistically significant one was eligibility for additional pay ($p = .004$). In other words, it is safe to generalize that participants who knew they were eligible to receive additional pay based on their summative performance ($n = 19$) rated the use of student growth percentiles 1.888 points higher than those who were not or did not know if they were eligible ($n = 130$). Interestingly, the coefficient for participation in the Pay for Performance program indicates that teachers who knew that they were in the program ($n = 17$) rated the practice lower in terms of propriety in relation to those who were not in the program or did not know ($n = 132$). While not statistically significant, it is also noteworthy that teachers with more than 5 years worth of experience, tended to rate the practice .569 points lower in terms of propriety than their more inexperienced counterparts; this is especially interesting because the significance level is only slightly

above the .05 threshold. As noted in Table 23 in Appendix A, the composite score descriptive statistics disaggregated by independent variables show a similar difference in means. Also apparent in Table 23 is an interesting difference in means of teachers in the 6+ years of experience category. Specifically, the mean score of teachers with 20 or more years experience (5.97) rated the use of student growth percentiles as lower than teachers with 6-20 years (6.51). In other words, the more experienced the teacher group was, the lower it rated the use of student growth percentiles in evaluation based on the construct of propriety. Like the variable of experience, previous evaluation experience also was not statistically significant ($p = .460$); however, the mean propriety composite score increased by a noticeable difference for those teachers who had been previously evaluated with student growth percentiles. Specifically, the mean for teachers previously evaluated with student growth percentiles was 7.04 while the mean for those who were not or did not know was 6.21 (see Table 23).

The factor with the lowest B value was professional affiliation. Specifically, teachers affiliated with the NEA or AFT had just a slightly smaller difference in composite score mean ($B = -.029$) as compared to their peers who reported not being a member of such a professional organization. This is noteworthy as many of the warnings against using tools such as student growth percentiles have been written and published by teacher unions and professional organizations.

To gain greater insight into the participants' perceptions of the use of student growth percentiles as it relates to propriety, they were offered an open-ended response option to make comments. Twenty-three percent of the participants ($n = 37$) volunteered some sort of response related to propriety. Propriety focuses on whether an evaluation

practice is legally defensible, ethically centered, and keenly focused on the welfare of the employee (Gullickson, 2009; Sanders, 1997). The themes that emerged from analysis of the comments included: the effect on instruction, parity among teachers, and student/parent accountability.

Effect on Instruction

The first theme to emerge form the comments was the effect that using student growth percentiles in teacher evaluation has on classroom instruction. This theme is included under the umbrella of propriety because of its close relation to the Service Orientation standard (Standard P1) which states that personnel evaluations should promote sound education of all students, fulfillment of institutional missions, and effective performance of job responsibilities (Gullickson, 2009). Four participants identified the positive potential of using student growth percentiles in evaluation as it relates to informing instruction, an occurrence that potentially reflects why the mean statement that using student growth percentiles aids in providing a sound education (2.30) was the 6th highest among all surveyed items (see Table 21 and Table 22 in Appendix A). While two of the responses were vaguely positive, the other two specifically noted that a "growth mindset" with instruction is a healthy approach and is supported by student growth percentiles. The fears of using student growth percentiles in evaluation far overshadowed the feeling of potential expressed above, with specific attention paid to how it may be a threat to a sound education for all students. Three participants expressed specific concern about how the practice encourages teachers to "teach to the test." Two other participants expressed concern that a focus on student growth percentiles will encourage the teaching of lower level thinking skills with Participant 13 noting:

> I feel as if teachers should be held accountable for student growth in his or her classrooms. However, I strongly believe this will negatively impact some teachers' instruction. Teachers will force students to memorize information and will retract from using best practices because of a push for high test scores. This will impact the school setting substantially.

In addition, three other participants noted the implementation of student growth percentiles in their evaluation took away time that they felt was better used in other ways. For example, Participant 64 noted that "a great amount of instructional time is now lost due to the pre-testing and post-testing we are now required to do" while Participant 31 stated, "I am a teacher who has always maintained 88% passing and above on all SOL testing subjects I have taught and am now being weighed down by paperwork and data collection even though I have always exceeded at my job."

Parity Among Teachers

The second theme that emerged from the open-ended responses was parity among teachers, a theme also related to the Service Orientation standard (Standard P1). Two participants expressed concern that only certain teachers are evaluated with student growth percentiles, implying that the learning experiences students receive from teachers who are not evaluated in such a manner can be different. For example, Participant 67 articulated, "Teachers in SOL grades are held to a different standard than those in non SOL grades/subjects. This makes the evaluation process uneven and unfair when assessing teachers." Two other participants, who were not only evaluated with student growth percentiles but also eligible for additional pay, advocated for their peers by noting that it is not fair for only a few teachers to be able to receive a stipend based on a judgment of performance using student growth percentiles.

Student/Parent Accountability

The final theme that emerged in regards to propriety was student/parent accountability. This theme is related to the Comprehensive Evaluation standard (Standard P5) because the comments alluded to the difficulty an evaluator may have "identify[ing] strengths and areas for growth" (Gullickson 2009) as much of the locus of control is out of the teacher's hands. Participant 23 noted:

> I find it interesting that lawyers and doctors are paid regardless if they win a case or lose a patient respectively. However, all burden is on the teacher without any consequences to students who do not give their all. You can lead a horse to water…

Participant 17 was more emotional in stating:

> I don't agree with this issue. I can't make the student study or do homework. There is no accountability for either the parent or the student. Why does it always fall on the teacher? In the real world you fire not working employees. To[o] bad you can't do that with students.

Participant 18 provided a case study of sorts to illustrate a similar point:

> I have to select one class [to focus on student growth], and I was told to select my collaborative class. I have 18 students in this class. Here is the breakdown:
>
> 1. Dad is in remission from cancer; however, kid is afraid dad will die.
> 2. Absent 43 days in addition to a time withdrawn and put into a court ordered placement. (as of 4/24)
> 3. Oldest of 5 siblings one of which has Asperger's
> 4. Expelled
> 5. Absent 18 days and little effort when here (as of 4/24)
> 6. 504
> 7. Possible undiagnosed Asperger's
> 8. Unmedicated ADD
> 9. Absent 20 days (As of 4/24)
> 10. LD
> 11. LD
> 12. LD and little effort
> 13. LD who really tries
> 14. Unmedicated ADHD
> 15. Unmedicated ADHD
> 16. LD
> 17. ED
>
> These children hold 40% of my evaluation in their hands. I have seen growth in most of them, but some have to be here to show growth. I only ask that those who

> have put this evaluation system into effect have the same challenges … and let's see how they feel. Teaching is not a surgical situation. We have to deal with children who are having to cope with adult situations and are not prepared for that. We have to deal with emotions, puberty, and social pressures that are really bad in some cases. Sometimes teaching academics takes a back seat to dealing with emotions. For example after the Connecticut school shooting, I lost a day of 'teaching' so I could respond to a child who told me he was afraid to be at school. What was more important? Dealing with that fear! At that point, I couldn't care about academics or my evaluation, or anything except those children's emotional state and having them feel as safe in their school as I possibly could. I dare any business person to be evaluated on this scale.

In all, 25 of the participants provided comments centered on the issue of the student and family having an undue amount of control in a teacher's evaluation, thus, making it difficult to truly assess his/her strengths and weaknesses. This may help explain why the mean for the statement regarding the ability to identify a teacher's strengths and weaknesses while using student growth percentiles (1.97) was the 6th lowest in the study (see Table 20 and Table 22 in Appendix A). Participant 60 summarized the comments as he/she noted, "two-thirds of the equation (student and parent) for educational success are not considered" and that "using student growth percentiles may be made legal by the state legislature, but it is not ethical, and has nothing to do with the welfare of the evaluatee."

Research Question 2: Utility

The second research question asks: What are teachers' perceptions of the use of student growth percentiles in terms of utility? To address this question, participants rated their level of agreement on a Likert scale (1 = *Strong Disagreement*, 2= *Disagreement*, 3 = *Agreement*, 4= *Strong Agreement*) regarding statements about the use of student growth percentiles as it relates to utility. Seven statements were categorized under issues of utility.

For the seven items, a composite score for the overarching category of utility was calculated. As noted in Table 29 (see Appendix) Cronbach's alpha had a value of .825. With a potential minimum of 6 (one question had N/A as an option) and a potential maximum of 28, the composite score mean was 14.26. Assuming a score of 15.5 to represent neutrality, the composite score indicates a slightly negative attitude toward the use of student growth percentiles in evaluation as it relates to utility. Participants' responses by individual item are summarized in Table 12.

Table 12

Teachers' Perceptions of the Use of Student Growth Percentiles as it Relates to Utility

Factor	N	Minimum	Maximum	Mean	Std. Dev.
Aids in retention of good teachers	149	1	4	2.17	.942
Reinforces good instruction practice	148	1	4	2.36	.934
Promotes sense of professionalism	146	1	4	2.21	.946
Fosters sense of collegiality	148	1	4	1.89	.907
Provides clear expectations	141	1	4	2.23	.981
Timely and functional reporting	64	1	4	2.40	1.303
Evaluator qualifications	143	1	4	2.73	1.048

Assuming a score of 2.5 indicates a balance of agreement and disagreement, it is noteworthy that all but one of the items studied had a mean below this score. The one item that scored above 2.5 was the statement, "My evaluator is qualified and capable of interpreting student growth percentiles data in my evaluation" with a score of 2.73. As seen in Table 21 and Table 22 in Appendix A, this is the second highest mean for all of

the individual statements, and the highest one for questions when "N/A" was not an option. Interestingly, "Timely and Functional Reporting" scored high relative to other factors; its means core of 2.40 represented the fourth highest among all survey questions. What makes this score stand out is that the timetable for Virginia's SOL testing is one where the actual testing does not occur until early or mid May, leaving little time to generate the necessary reports prior to the close of most schools in June, when teacher summative evaluation are due. This could indicate either that the schools that the participants taught in received reports in a very prompt manner, or despite not having the reports promptly, when compared to the other survey questions, this was a less disagreeable statement. The factor that scored lowest was fostering a sense of collegiality with a mean score of 1.89. As Table 20 and Table 22 in Appendix A also details, this is the second lowest mean score for any individual statement in the survey.

To determine the predictive value of key independent variables, a regression analysis was run on the utility composite scores. Coding for the independent variables was the same as the coding for the regression analysis for propriety and reflected in Table 8. As indicated in Table 13, the R Square value of .195 in the model summary of the regression analysis indicates that 19.5% of the variance in composite score means can be explained using the model.

Table 13

Model Summary of Utility Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.442	.195	.167	4.649

In addition, the ANOVA for this particular regression model reflected in Table 14 reveals a significance of .000, indicating an excellent model in terms of fit.

Table 14

ANOVA of Propriety Regression Model

Model	Sum of Squares	df	Mean Square	F	Sig.
1	750.643	5	150.129	6.947	.000
	3090.149	143	21.609		
	3840.792	148			

Considering the strong significance level of the regression model, it is valuable to examine the coefficients detailed in Table 15.

Table 15

Regression Table for Utility Composite Scores

	B	SE B	β	t	Sig.
(Constant)	25.214	3.502		7.199	.000
Experience	-2.041	.580	-.271	-3.518	.001
Professional Affiliation	.540	.782	.053	.690	.491
Previous Evaluation Experience	-1.650	.867	-.156	-1.904	.059
Eligibility for Additional Pay	-3.861	1.224	-.254	-3.154	.002
Participation in Pay for Performance Program	2.104	1.223	.132	1.720	.088

Two variables provide particularly strong predictive value. The years of experience category, with a statistical significance of .001, shows that it is safe to generalize that the mean score for teachers with 6+ years of experience (n = 117) is 2.041 points lower on average that teachers with 5 or fewer years (n = 32) in regards to utility. In addition, with a significance level of .002, the variable for eligibility to receive addition pay also proved predictive. Specifically, it is safe to generalize that those teachers who were eligible for additional pay scored the practice of using student growth percentiles in teacher evaluation 3.861 points higher than their counterparts. As noted in Table 23 in Appendix A, the composite score descriptive statistics disaggregated by independent variables show a similar difference in means. Strikingly, the trend is

reversed for teachers reporting that their school is a part of the Pay for Performance program; while not statistically significant ($p = .088$), it is worth noting that teachers reporting participation in the Pay for Performance program ($n = 17$) have a mean that is on average 2.104 points lower than those who are not or do not know ($n = 132$). Another variable worth noting was previous evaluation experience with a significance level above .05 ($p = .059$). Specifically, those teachers reporting having a summative evaluation using student growth percentiles the previous year scored the practice 1.650 points higher on average in terms of utility than their counterparts. The trend toward rating the practice of using student growth percentiles higher if teachers had been previously evaluated in such a manner is also reflected in the increase in means in Table 23. The mean for teachers who had previously been evaluated with student growth percentiles was 15.96 while the one for those who were not or did not know if student growth percentiles had been used in their evaluation was 13.32.

To color the findings reflected in the statistics regarding participants' perceptions of the use of student growth percentiles as it relates to utility, and open-ended response option was made available to participants make comments. Coding of all of the comments demonstrated that 17% of the participants ($n = 26$) volunteered some sort of response related to utility. While utility focuses on whether an evaluation is sufficiently "informative, timely, and influential" (Gullickson, 2009, p. 69), the themes that emerged from analysis of the comments included: collegiality and professionalism, evaluator's role, and timeliness.

Collegiality and Professionalism

The first theme to emerge was collegiality and professionalism, a topic that falls under the umbrella of the Constructive Orientation standard (Standard U1) as stated by Gullickson (2009). All seven of the participants that commented on the theme had a negative attitude, one summarized by Participant 60 who wrote that using student growth percentiles in evaluation "would create an atmosphere of fear, intimidation and anxiety" and Participant 19 who stated, "It breeds an aura of hostility and desperation in a school." Three participants expressed that teachers in their school no longer collaborate like they once did because of the increased feeling of competition. In addition, Participant 13 predicted that teachers would begin "fighting over specific students" when class rolls are created.

Evaluator's Role

The next theme to emerge was the evaluator's role in evaluation, one directly linked to the Evaluator Qualifications standard (Standard U3) that states an evaluation system should be developed, implemented, and managed by people with the necessary skills, training, and authority (Gullickson, 2009). Nineteen participants contributed comments centered on the theme of an evaluator's role with none of them having a positive attitude. This is a noteworthy occurrence since the mean score for the statement regarding evaluator qualifications (2.73) was the second highest in the study (see Table 21 and Table 22 in Appendix A). While most of the comments were general in nature regarding evaluators, three participants expressed specific concern that administrators did not truly understand student growth percentiles. For example, Participant 56 articulated:

> The current principal and administrator corps owns fundamental misunderstandings of what student growth percentiles are, how they reflect (or

> don't) the work of the classroom teacher, and how they can be used to evaluate a teacher's performance. There is little to no room for discussion of how or why certain results exist; simply identifying the numbers and recording them on the evaluation document is sufficient to the process of judging teachers by SGPs (or other growth models). To boot, evaluators often have little to no understanding of the difference between criterion-referenced evaluation and norm-referenced thinking; in many places SGPs are guaranteeing winners and losers, which destroys any value they held to begin with.

Two participants noted that having principals with no teaching experience made the analysis of student learning gains a flawed system. Perhaps more disconcerting than the perceived lack of administrative competence in using student growth percentiles was the feeling that some administrators may manipulate the system for various purposes other than providing a fair teacher evaluation. Participant 47 wrote, "A score can be manipulated [to] show whatever data is needed/ desired/wanted for whatever purpose the evaluator/s need it to show." Participant 3 provided an example: "… there are administrators who unfortunately seem to set those struggling teachers up for failure by compiling a list of challenging students."

Timeliness

The final theme to emerge from comments centered on utility was timeliness. Six participants made comments regarding timeliness, with half of those noting that the teacher evaluation process in general in their schools was not done in a timely manner. The other three specifically noted that the use of student growth percentiles was not timely. Participant 11 noted, "In order to be of benefit for a yearly evaluation, the SGPs should be issued before the end of the school year." Not having student growth percentiles at the end of a current academic year was also alluded to by Participant 76 who stated, "I would love to know how my students performed on the SOL in June, right after they take it, not in September" and Participant 46 who argued, "…using past SGP to

do current evaluations could be confusing and not necessarily timely." These comments are particularly interesting because like the statement regarding evaluator qualifications, the statement regarding timely and function reporting had a relatively high score of 2.40 represented the fourth highest mean among all survey questions (see Table 21 and Table 22 in Appendix A).

Research Question 3: Accuracy

The final research question asks: What are teachers' perceptions of the use of student growth percentiles in terms of accuracy? The same survey techniques used for the first two research questions were used to address this question. Specifically, participants rated their level of agreement on a Likert scale (1 = *Strong Disagreement*, 2= *Disagreement*, 3 = *Agreement*, 4= *Strong Agreement*) regarding statements about the use of student growth percentiles as it relates to accuracy. Eleven statements were categorized under issues of accuracy.

For the 11 items, a composite score for the overarching category of accuracy was calculated. As noted in Table 29 (see Appendix), a value of .836 was calculated for Cronbach's alpha. With a potential minimum of 9 (two statements had N/A as an option) and a potential maximum of 44, the composite score mean was 21.39. Assuming a score of 26.5 to represent neutrality, the composite score indicates an overall negative attitude toward the use of student growth percentiles in evaluation as it relates to accuracy. Participants' responses by individual items are summarized in Table 16.

Table 16

Teachers' Perception of the Use of Student Growth Percentiles as it Relates to Accuracy

Factor	N	Minimum	Maximum	Mean	Std. Dev.
Evaluation judgments with student growth percentiles was similar to classroom observations.	63	1	4	2.66	1.591
Summative score with student growth percentiles was similar to previous years.	63	1	4	2.85	1.560
Allows for adequate accounting of influence of previous teachers.	148	1	4	2.18	.864
Allows for adequate accounting of influence of support personnel.	145	1	4	1.99	.861
Allows for adequate accounting of school scheduling.	146	1	4	1.96	.917
Allows for adequate accounting of school resources.	144	1	4	1.94	.868
Allows for adequate accounting of student characteristics.	144	1	4	1.96	.982
Allows for adequate accounting of community and cultural influences.	145	1	4	1.83	.890
Standardized tests adequately measure what students should learn.	145	1	4	2.10	.964
Reliability	145	1	4	2.12	.887
Student growth percentiles make mathematical sense.	145	1	4	2.24	.880

Using a score of 2.5 to indicate a balance of agreement and disagreement, two of the statements had means that were higher, indicating general agreement. Specifically, the means for "My evaluator's judgments of my performance using student growth percentiles was similar to his/her judgments of my performance based on classroom observations" and "My summative (final) rating in 2011-2012 was similar to summative ratings I have received in the past" scored a 2.66 and 2.85. As noted in Table 21 and Table 22 in Appendix A, these represent the highest and third highest means of the individual statements. It is also worth noting that both of these statements had "N/A" as an option, reducing the number of scored responses (n = 63) for each statement. The standard deviations for each statement (1.591 and 1.560) were the two highest on the survey, indicating a large amount of variance in the responses relative to other statements.

Of the 11 statements regarding accuracy, 5 of them had a mean below 2.0, indicating a strong negative attitude toward each statement (see Table 20 and Table 22 in Appendix A). In addition each statement centers on the influence of outside factors on student achievement. Ranking lowest of all the factors studied was "Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of community and cultural beliefs and practices on student achievement" with a mean of 1.83. The mean for "Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of school resources (e.g., curricula, textbooks, furniture, etc.) on student achievement" scored the third lowest at 1.94. The means for statements regarding the adequate accounting of the influences of school

scheduling and student characteristics were both 1.96. Finally, the statement regarding the adequate accounting of the influence of support personnel was 1.99.

To determine the predictive value of key independent variables, a regression analysis was run on the accuracy composite scores. Coding for the independent variables was the same as the coding for the regression analysis for propriety and is reflected in Table 8. As indicated in Table 17, the R Square value of .103 in the model summary of the regression analysis indicates that 10.3 % of the variance in composite score means can be explained using the model.

Table 17

Model Summary of Accuracy Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.321	.103	.071	7.114

In addition, Table 18 confirms the model to be a good fit with a significance of .008.

Table 18

ANOVA of Accuracy Regression Model

Model	Sum of Squares	df	Mean Square	F	Sig.
1	829.233	5	165.847	3.277	.008
	7236.189	143	50.603		
	8065.423	148			

Considering the significance level, it is helpful to examine the coefficients of each independent variable detailed in Table 19.

Table 19

Regression Table for Accuracy Composite Scores

	B	SE B	β	t	Sig.
(Constant)	32.992	5.360		6.156	.000
Experience	-2.150	.888	-.197	-2.421	.017
Professional Affiliation	.632	1.196	.043	.528	.598
Previous Evaluation Experience	-1.532	1.326	-.100	-1.155	.250
Eligibility for Additional Pay	-4.281	1.873	-.194	-2.285	.024
Participation in Pay for Performance Program	2.266	1.872	.098	1.211	.228

As was the case when analyzing the utility and propriety composite scores, the variable of experience was the most statistically significant predictor of an average change in means ($p = .017$). With a significance level like this that is below the .05 threshold for 95% confidence interval, it is safe to generalize that teachers with 6+ years of experience scored the practice of using student growth percentiles 2.15 points lower on average that teachers with 0-5 years. As noted in Table 23 in Appendix A, the composite score descriptive statistics disaggregated by independent variables show a similar difference in means. As was the case with the previous two regressions, eligibility to receive pay was also a statistically significant variable ($p = .024$) and provided the largest change in average means ($B = -4.281$). While we can safely assert that teachers who report being eligible to receive additional pay score the practice of using student growth percentile in evaluation 4.281 point higher on average, it does come with its caveats. Once again, the number of teachers who reported being eligible to receive additional pay was small ($n = 19$) when compared to those who were not, did not know, or did not teach the previous year ($n = 130$). In addition, there was a similar trend as the one described in

the regression for utility composite scores: teachers who reported being a part of the Pay for Performance program rated the practice 2.266 points lower on average. Once again, it is important to note the number of teachers reporting being a part of the program (n = 17) as opposed to those who were not, did not know, or did not teach the year before (n = 132). Finally, while the coefficient for previous evaluation experience was not statistically significant (p = .250), it is worth noting that the teachers who had previously been evaluated with student growth percentiles rated the practices higher than those who had not been evaluated with them or did not know, a trend that was also reflected in the utility and propriety composite scores. Specifically, the mean for teachers who had been previously evaluated with student growth percentiles was 23.08 while the mean for their counterparts was 20.46 (see Table 23).

To provide context to the findings from the closed questions on the survey, participants were provided the opportunity to respond to an open-ended question regarding the accuracy of using student growth percentiles in teacher evaluation. Coding of all of the comments demonstrated that 45% of the participants (n = 68) volunteered some sort of response related to accuracy. While accuracy focuses on whether an evaluation is "technically adequate and complete to produce sound information appropriate for the purpose of making sound judgments" (Gullickson, 2009, p. 115) the themes that emerged from analysis of the comments included: support personnel and resources, previous teachers, scheduling, student characteristics and demographics, and the calculation of student growth percentiles.

Support Personnel and Resource

Five participants provided comments that centered on the theme of support personnel and resources, a topic considered a part of the Analysis of Context standard (Standard A3). Participants 8 and 80 cited a lack of equity in term of how support personnel are allocated within their schools. Specifically, Participant 8 noted, "Class size and support personnel are not evenly distributed throughout grade levels and the evaluation progress remains the same way regardless" while Participant 80 wrote:

> In our school some grades receive support personnel to help with weaker students while others do not get any, so how can you compare when some years they had extra support while others they did not?

The other two responses expressed concern that resources were not equitably distributed from district to district. For example, Participant 13 wrote:

> I agree that as a teacher, I should base my instruction off of data showing where each student is, what they have mastered and what they have not mastered. I do believe that this data can assist us in making a judgment based on students--what classes they are in, what resources they need, etc. What I don't understand is how this correctly reveals the background students have such as where they are from, how many additional resources were needed or still are needed, and opportunities in the county (textbooks, technology) … Our county does not receive more or less money and resources that I am aware of because we have lower or higher student growth percentiles.

Participant 15 was more specific, stating:

> Our county does not have 1 reading specialist. Is it fair to compare my students to counties who have 1 (+) specialist per building? Our county provides instructional support (PALS, Title I, etc), but there are not support personnel for 6th/7th grade reading. We do have Algebra Readiness support for math at those levels. Is that fair? Reading needs are not miraculously 'resolved' at the 6th grade level. When the state of Virginia begins funding all counties with the same money per child--maybe then we could compare SGPs.

These perspectives dovetail with the mean scores of the corresponding statements regarding the influence of school resources (mean = 1.94) and support personnel (mean =

1.99), both of which were below 2.0, a score indicating a strong negative attitude toward each statement.

Previous Teachers

The next theme to surface, albeit by just three participants, was the influence of previous teachers, also an element of the Analysis of Context standard (Standard A3). Participants 46 and 54 emphasized that whether it be for the benefit or detriment of a student, student growth percentiles did allow room for an evaluator to account for the influence of previous teachers. Participant 9 provided a more vitriolic stance:

> S.O.L. testing does not begin until grade 3, but the kindergarten, first, and second grade learning is included in 3rd grade S.O.L. testing. If the teachers in those grades don't take their teaching seriously because they aren't tested (or for any other reason), they don't suffer, but put the 3rd grade teacher in the position of trying to catch up several years of teaching. (I have heard teachers of untested grades talking about how glad they are that they don't have the expectations that S.O.L. testing grades have.) So while mathematically student growth percentiles may seem to make sense, it leaves out the other factors that S.O.L. testing grade teachers must face and attempt to overcome.

The lack of comments regarding the influence of previous teachers may be a reflection of teachers' belief that while student growth percentiles may not adequately reflect their influence as evidenced in the mean score of 2.18, it is not as problematic as other factors because the score also is safely in the middle range of rank scores (see Table 22 in Appendix A).

Scheduling

The third theme to emerge from the open ended responses was scheduling, another theme directly aligned with the Analysis of Context standard (Standard A3). As evidenced by the fact that 16 different participants provided comments on the issue and each expressed concern, it appears this is a topic worth considerable attention. Class size

represented one of the concerns as Participant 46 expressed how schools with smaller teacher to pupil ratios would likely perform better. Participant 8 echoed how helpful smaller class sizes are but shared that there can be great variation even within the same school. Another concern that was raised was the length of the instructional period. For example, Participant 16 recounted:

> Last year my students' scored well with an 89.5 pass rate. I had an 82 minute period to teach those students. This year my periods are 60 minutes due to more time being given for Reading instruction. However, I am being expected to produce the same or better results with a 25% reduction in instructional time.

The scheduling factor that received the most attention, however, was class grouping as it was the topic of the remaining 13 responses. The majority of the comments expressed concern that teachers who teach "at risk" student may not have their effectiveness reflected in student growth percentiles. For example, Participant 21 shared:

> I think that it is not 100% fair to evaluate students from year to year when student groups can vary, especially if you have gifted versus a low group. If I have gifted one year and low the next, it will make me look like a worse teacher. And the same can be said for the opposite; if I have low one year and gifted the next, it will make me perhaps look better than I am. Also, it doesn't seem right that a teacher with a gifted class might receive more rewards for their work over a teacher with a low class who may be working harder with their students, but isn't achieving the same results.

This comment was representative of all but one participant who astutely pointed out that the opposite is true when using student growth percentiles. Specifically, Participant 26 demonstrated:

> If I am evaluated on my student's growth throughout the year, and not their SOL score, then I will feel fairly evaluated. My students' SOL scores should not be compared to the SOL scores of another teacher, because we teach based on 'grouping.' A lower group may score lower on the SOL, but have the most growth throughout the year.

It is difficult to ascertain from the comments if the discrepancy in attitudes is a matter of not fully understanding student growth percentiles and how they are calculated or if there is a genuine belief that students below grade level can not demonstrate academic progress as easily as students on grade level or above. Regardless of the cause, the comments reflect the same negative attitude regarding whether the use of student growth percentiles allows for the adequate accounting of scheduling (mean = 1.96) represented in the closed responses.

Student Characteristics

The last theme to emerge associated with the Analysis of Context standard (Standard A3) was the influence of student characteristics. This was clearly a topic of interest and concern for participants as 37 voluntarily provided comments. In terms of observable characteristics, the one that was consistently identified was socio-economic status. For example, Participant 15 stated, "I work in a minimally funded county, and it is discouraging to know that my students (and my own children) will be compared to students in NOVA." Participant 73 expounded on this stance:

> Using growth percentiles (test scores) in order to evaluate a teacher's work (performance) is neither ethical nor productive in attempting to ascertain a teacher's impact on student growth during a school year. Especially in impoverished areas, so much of the impact a teacher has on a student is developed in the heart of the student (not measureable), not the tip end of the pencil he is using.

Another characteristic that surfaced was student behavior. As Participant 74 articulated:

> Using student growth percentiles in teacher evaluations does not take into consideration the behavioral issues in the classroom. Even if the teacher has no issues with classroom management, there are always [a] couple [of] students who are always playing around and becoming the source of distraction for the rest of the class.

Similarly, two participants specifically noted that students who are habitually absent can skew results. Additionally, participants noted the influence that unpredictable emotional events can have on students such as: divorce (Participant 9), mental illness (Participants 9 and 13), and mood swings (Participant 49). The student characteristic that dominated the responses, however, was student motivation. Participant 47 articulated, "Motivation to learn cannot be LARGELY instilled by teachers." Participant 2 went so far as to suggest a change in the school and testing calendar to address student motivation concerns:

> A teacher can teach, practice, review, and review again and again and some students still will not pass the test. Give the students more incentive. Have all students take the tests at the end of April and those passing the test are released for summer vacation. Have those failing the test stay in school and take part in remediation skills for 3 weeks, then retake test. I know with this incentive all students would do well on tests. For some, all it takes is the right motivation. If we tried this approach we would know that we had given our all to all students and if they had not passed the test it was not because they had not been taught the material. There are practical ways of getting better results on tests in the state of Virginia.

The volume of comments may help explain the low means generated when participants were asked if student growth profiles adequately allow for the influence of students' characteristics (mean = 1.96) and community and cultural circumstances (mean = 1.83).

Test Characteristics

The next theme to emerge from the comments, impressions of the SOL tests themselves, was associated with the Defensible Information standard (Standard A5). One concern that arose was that the reading level of the tests may prohibit students from demonstrating growth in a given subject. As Participant 41 noted, "As long as children at a 2nd grade reading or math level are tested at 4th grade level, then their scores do not accurately reflect what they have learned." Three other comments also echoed these sentiments, expressing concern that the appearance of no growth may be the result of the

inaccessibility of the test due to the reading level. The other major concerns regarding the SOL tests was the introduction of technology enhanced items that were introduced during the same time period the student growth percentiles were. As Participant 56 noted:

> In Virginia, student growth percentiles are based on tests that include Technology Enhanced Items (TEIs). TEIs are scored in an all-or-nothing manner; on 'hot spot' questions, for example, a student who identifies 11 or 12 choices correctly receives the same score as a student who identifies 0 of 12. As long as Virginia continues to include ridiculous questions like these, results will be inaccurate, as will any connection made between test scores and teacher performance.

The emergence of these themes may reflect a reason for why the statement regarding how well the standardized tests used measure what they should for student growth percentiles had the 8th lowest mean (mean = 2.10) among the closed survey items (see Table 22 in Appendix A).

Student Growth Percentile Calculations

The final theme to emerge from the comments provided by participants centered on the validity and reliability of the student growth percentile calculations, a theme associated with the Defensible Information standard (Standard 5), Reliable Information standard (Standard A6) and Analysis of Information standard (Standard A9). One criticism centered on the comparability of students' scores. While one participant (Participant 56) argued that it is unfair to compare student scores from impoverished areas of the state to student scores form more affluent areas, three other participants stated that student academic growth should be measured against the individual's past performance, not the performance of anyone else. Another concern that emerged was that there were not enough data points to ensure confidence in student growth percentiles.

Participant 39, while vehement in tone, summarized the concerns expressed by 6 other participants:

> Student growth percentiles are just one minute second in a student's educational career. How can a SOUND JUDGEMENT be made based on a second? This … is a bunch of bovine offal. Student growth percentiles are based on one test, one day out of the 180 days that I work with individuals, not machines. There are so many unmeasurable factors that encompass those individuals that student percentiles are mutely unfair.

The final major concern regarding the calculations of student growth percentiles was the omission of students with advanced pass scores. Participant 9 succinctly stated, "I don't understand why students who make scores of 500 or more are not considered in the scoring of student growth." While these concerns reflect the relatively low score the reliability statement earned (mean = 2.12), it is interesting to note that the statement regarding how student growth models make mathematical sense, while still below the score of 2.5 that would reflect neutrality, scored the 7th highest when ranked ordered (mean = 2.24), raising the question of whether teachers understand student growth percentiles.

Summary

Participants in the study had a negative attitude toward the use of student growth percentiles in teacher evaluation in all three of the domains analyzed, as evidenced by composite scores that all fell below the identified scores for neutrality. The composite scores for propriety, utility, and accuracy were 6.5, 14.26, and 21.39 respectively; the scores indicating a balance of agreement and disagreement for each domain were 7.5, 15.5, and 26.5. Of these three domains, accuracy was rated lowest as evidenced by the fact that it came the closest of the three domains to having a mean that is a full standard deviation from the indicated neutral score (see Table 23).

In the area of propriety, there was a primary concern regarding how using student growth percentiles allows for a comprehensive evaluation. Specifically, the mean for the statement, "Using student growth percentiles in teacher evaluation allows for the strengths and weaknesses of a teacher to be shown in the evaluation," was the lowest among the three at 1.97, the 6th lowest among all statements on the survey (see Table 20). Additional concerns were noted in the teachers' comments that centered on the potentially negative effect on instruction, the small number of teachers who were eligible to be evaluated with student growth percentiles, and the amount of control that students have on a teacher's evaluation.

Table 20

Rank Order Statements With Means Below 2.0

Factor	N	Min.	Max.	Mean	Std. Dev.
Allows for adequate accounting of community and cultural influences.	145	1	4	1.83	.890
Fosters sense of collegiality	148	1	4	1.89	.907
Allows for adequate accounting of school resources.	144	1	4	1.94	.868
Allows for adequate accounting of school scheduling.	146	1	4	1.96	.917
Allows for adequate accounting of student characteristics.	144	1	4	1.96	.982
Reflects teacher's strengths and weaknesses.	148	1	4	1.97	.903
Allows for adequate accounting of support personnel.	145	1	4	1.99	.861

In the area of utility, there existed the only positive attitude that can be confidently gleaned from the study. The statement, "My evaluator is qualified and

capable of interpreting student growth percentiles data in my evaluation," had a mean of 2.73, the second highest among all statements on the survey (see Table 21). It is the only statement that can be confidently considered positive, however, because the other three statements with means above 2.5 (the identified score for neutrality) had a large number of participants mark "N/A" as a response. There was concern about the effect that using student growth percentiles would have on collegiality in a school as evidenced by a mean of 1.89 for the statement, the lowest mean among statements in the utility domain and the second lowest of all statements in the study (see Table 20). The concern about collegiality was also reflected in the open-ended responses, as was concern regarding how qualified supervisors were in terms of using and understanding student growth percentiles, as well as the lack of timeliness when they were used.

Table 21

Rank Order of Statements With Means Above 2.5

Factor	N	Min.	Max.	Mean	Std. Dev.
Summative score with student growth percentiles was similar to previous years.	63	1	4	2.85	1.560
Evaluator qualifications	143	1	4	2.73	1.048
Evaluation judgments with student growth percentiles was similar to classroom observations.	63	1	4	2.66	1.591

As the lowest rated category in terms of composite score as well as the one with the most items analyzed, the area of accuracy produced the most low rated items regarding the use of student growth percentiles. Five of the statements had means below 2.0, indicating strong disagreement among participants (see Table 20). The lowest mean

for not only the accuracy domain but also the entire study was for the statement, "Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of community and cultural beliefs and practices on student achievement" (mean = 1.83). Also scoring below 2.0 were statements centering on the influence of the following factors: school resources (mean = 1.94), scheduling (mean = 1.96), student characteristics (mean = 1.96), and support personnel (mean = 1.99). In addition, specific concerns were noted in the open-ended responses about support personnel and resources, previous teachers, scheduling, student characteristics, test characteristics, and student growth percentile calculations.

Regression analyses of each domain identified some predictive value for independent variables. The variable for experience was most predominant across the domains as teachers with 6+ years' worth of experience generally had a more negative attitude toward the use of student growth percentiles in terms of propriety, utility, and accuracy. In addition, the variable was statistically significant for utility ($p = .001$) and accuracy ($p = .017$). The other variable with a strong predictive value was eligibility to receive additional pay as those respondents reporting being eligible rated all three domains higher than their counterparts who either were not eligible or did not know if they were eligible. In each category, those who reported being eligible for additional pay rated the practice higher than their counterparts. Despite the fact that eligibility to receive pay was a statistically significant coefficient in all of the domains, there is reason to question the results because in all three domains participants who cited being a part of the Virginia Pay for Performance program all rated the use of student growth percentiles lower than their counterparts. Finally, while previous evaluation experience was not a

statistically significant coefficient in the regression analyses for each composite score, the means for each composite score were higher for teachers who had previously been evaluated with student growth percentiles, indicating that going through the process of an evaluation with student growth percentiles may help create a more positive attitude toward the practice.

CHAPTER 5

SUMMARY AND CONCLUSIONS

Purpose

As reports of the national educational system more and more paint a negative picture of schools and the students they produce (Associated Press, December 7, 2010; Klein, 2011; NAEP, 2011), policy makers have begun to reexamine teacher evaluation as part of their educational reform initiatives (Newton, Darling-Hammond, Haertel, & Thomas, 2010). In examining teacher evaluation systems, the U.S. Department of Education (2010) has made it a priority to promote the inclusion of student academic gains by requiring it as a condition of NCLB waivers and grant programs. The Commonwealth of Virginia has implemented a system where student academic gains currently count for 40% of a teacher's evaluation. One way this 40% may be determined is by "other academic indicators" such as benchmarks tests, goal setting, and other more traditional forms of measuring student achievement. The other option is to have 20% determined by "other academic indicators" and the remaining 20% determined by student growth percentiles (VDOE, 2011b). Despite the growing momentum to use student

learning measures in teacher evaluation in states such as Virginia, there has been a tremendous amount of concern expressed regarding the practice, especially when standardized tests are used as the basis for such calculations (VEA, 2011). Considering these concerns and the limited research on teachers' perceptions regarding this practice, the purpose of this study was to determine teachers' perceptions of the use of student growth percentiles in the teacher evaluation process with a focus on three key areas: propriety, utility, and accuracy.

Research Questions

The following research questions guided this study:

1. What are teachers' perceptions of the use of student growth percentiles in teacher evaluation in terms of propriety?
2. What are teachers' perceptions of the use of student growth percentiles in terms of utility?
3. What are teachers' perceptions of the use of student growth percentiles in terms of accuracy?

Methodology

A survey was developed to assess teachers' attitudes about the use of student growth percentiles in evaluation using relevant standards from *The Personnel Evaluation Standards* (Gullickson, 2009). One of the populations selected for participation in this study was the group of teachers who were part of the Virginia Pay for Performance program in 2011-2012 and who returned to their same division the next year. School divisions that participated in this pilot program used student growth percentiles when available in teachers' evaluation. This population was chosen because of their experience

with the practice of using student growth percentiles in a high stakes environment where money was used to incentivize teacher performance. The other survey population included teachers in similar schools according to demographics and AMO status who may or may not have been evaluated using student growth percentiles and who may or may not have been eligible for additional pay like those who were in the Pay for Performance program. All selected participants received a pre-notification letter and later received the web-based survey via e-mail.

The survey included both open and closed ended questions in order to gain a rich assessment of teachers' beliefs while maintaining anonymity. Of the 399 surveys sent out, 150 completed ones were returned for a 38 per cent return rate. Using the information gathered through the survey, descriptive statistics were used to calculate the mean and standard deviation of individual items. In addition, composite scores were calculated for each of the main domains (propriety, utility, and accuracy) studied. To determine the predictive value of key independent variables, a regression analysis was run on the composite scores. Finally, the open-ended responses were analyzed to find emergent themes as they related to the three domains that anchor the research questions.

Summary of Findings

The study centered on three research questions. For each question, a composite score was calculated for the domain being studied (propriety, utility, or accuracy). Means and standard deviations were also calculated for individual items that fell under the umbrella of each domain. A regression analysis was also run on each composite score to identify variables with predictive values. Finally, open-ended responses were analyzed for each domain.

Propriety

The data collected and analyzed for the first research question, "What are teachers' perceptions of the use of student growth percentiles in teacher evaluation in terms of propriety?" revealed an overall negative attitude toward the practice. The composite score for the three statements, 6.50, was below the score that would indicate a balance of agreement and disagreement, 7.5. In other words, participants were not generally confident that using student growth percentiles is legally defensible, ethically centered, and keenly focused on the welfare of the employee (Gullickson, 2009; Sanders, 1997).

The highest mean rating was for the statement, "Using student growth percentiles in teacher evaluation promotes a sound education for students" at 2.30. While this was the highest mean for this particular domain, it was below 2.5, a score that would indicate neutrality (i.e., the average of the extremes of 1 and 4 on the Likert scale). In addition, comments from the open-ended responses expressed concern that using student growth percentiles may negatively affect instruction by "teaching to the test" and encouraging lower level thinking skills, concerns that have also been echoed in the literature on using student academic gains in evaluation (Braun, et al, 2010; Corcoran, 2010; Darling-Hammond, 2009; Darling-Hammond & Rustique-Forrester, 2005; Marshall, 2009; Misco, 2008; Mujis, 2005, Peterson, 2000; Ravitch, 2010).

Scoring lowest in the domain and 6th lowest on the entire survey was the statement, "Using student growth percentiles in teacher evaluation allows for the strengths and weaknesses of a teacher to be shown in the evaluation" at 1.97. The open-ended responses identified the need for teachers to respond to more than just students'

academic needs and the degree of agency that students and parents have as hurdles to identifying a teacher's strengths and weaknesses. The research on the topic has been particularly sure to note that teachers and principals often acknowledge the need to prioritize the teaching non-academic concerns (Borman & Kimball, 2005; Darling-Hammond, 2009; Harris & Sass, 2009; Stronge et al., 2007).

A regression analysis of the propriety composite scores revealed one statistically significant coefficient, eligibility to receive additional pay ($p = .004$). While the regression indicates that teachers who reported being eligible to receive additional pay rated using student growth percentile in evaluation 1.888 points higher than their counterparts, the results are questionable. The relatively small number of teachers identifying themselves as being eligible ($n = 17$) and a reverse trend for teachers who reported being part of the Virginia Pay for Performance program ($B = .432$) make this an area for further study more than an area for political and administrative action. While not statistically significant, the coefficient for experience ($p = .062$) indicated that teachers with 6+ years of experience rated the use of student growth percentiles lower than their less experienced counter parts ($B = -.569$). Interestingly, considering how the VEA (2010) and AFT (2010) have both published reports on the potential concerns of using student academic gains in evaluation, the coefficient for affiliation with a professional organization had the lowest B value ($B = .029$). This result encourages the conclusion that the negative attitude toward using student growth percentiles pervades among both members and non-members of a professional organization, and that professional organizations may not have much influence on teachers' opinions.

Utility

Similar to the results for the domain of propriety, the data collected and analyzed for the second research question, "What are teachers' perceptions of the use of student growth percentiles in terms of utility?" indicated an overall negative attitude. Specifically, the composite score for the 7 questions was 14.26, 1.24 points lower than 15.5, a score that would represent a balance of agreement and disagreement. To frame it in the context of the definition of utility, participants were not confident that using student growth percentiles in a teacher's evaluation is sufficiently "informative, timely, and influential" (Gullickson, 2011, p. 69).

The highest mean in the domain of utility and one of only three on the survey to score above a neutral ration of 2.5 was the one for the statement, "My evaluator is qualified and capable of interpreting student growth percentiles data in my evaluation" (mean = 2.73). This is a relatively surprising result as a healthy body of literature has noted that administrators may not be provided sufficient training in evaluation practices (Hess, 2007; National Governors Association, 2011), and many may not possess the necessary skills in teacher evaluation (Jacob & Lefgren, 2006; Kimball, 2002; McCaffery & Hamilton, 2007; Zimmerman & Deckert-Pelton, 2003). The next highest mean in the utility domain was for timely and functional reporting (mean = 2.40) which fell slightly below the neutral point on the Likert scale. This is also noteworthy because comments from the open-ended portions of the survey identified not having tests scores available at the end of the school year in which the standardized tests were taken as a concern. Webster and Mendro (1997) also noted a similar dilemma when using standardized tests as part of a teacher's evaluation.

The lowest mean in the domain for utility, and 2nd lowest in the study, was for the statement, "Using student growth percentiles in teacher evaluation helps foster a sense collegiality among staff members at a school" (mean = 1.89). The open-ended responses used words such as "hostility," "desperation," "fear," "anxiety," and "intimidation" when describing the effect on a school's atmosphere when student growth percentiles are used. Researchers have warned that such attitudes can surface in such a context (Baker, et al., 2010; Cowart & Myton, 1997; Schalock & Schalock, 1997; Rosenberg & Silva, 2012; Wise & Pease, 1983).

The regression analysis of the composite scores identified the same two coefficients, experience and eligibility for additional pay, as statistically significant ($p = .001$ and $p = .002$ respectively). Specifically, teachers with 6+ years of experience scored the practice of using student growth percentiles in evaluation in regards to utility 2.041 points out of 25.214 total points lower than their counterparts. While teachers who reported being eligible for additional pay rated the practice 3.861 points higher than their counterparts, the same questions regarding reliability about this particular variable arose that did in the first research question. While not statistically significant ($p = .059$), it is worth noting that teachers who reported having student growth percentiles as part of their evaluation the previous year rated the practice 1.650 point higher than those who did not or did not know. This finding may suggest that teachers might become more comfortable with the practice the more they are exposed to it.

Accuracy

The data gathered and analyzed for the final research question, "What are teachers' perceptions of the use of student growth percentiles in terms of accuracy?" was

the most striking of the three domains. The mean composite score of the 11 statements under the domain of accuracy was 21.39, the closest to a full standard deviation below the score that would indicate a balance of agreement and disagreement, 26.5 in this case, among the domains studied. In terms of the definition of accuracy, the composite score indicates that participants in general did not trust that using student growth percentiles in a teacher's evaluation was "technically adequate and complete to produce sound information appropriate for the purpose of making sound judgments" (Gullickson, 2009, p. 115).

Interestingly, two of the statements under the accuracy domain resulted in the highest and 3rd highest means in the entire study. Specifically, the mean for the statement, "My summative (final) rating in 2011-2012 was similar to summative ratings I have received in the past" was 2.85 while the mean for "My evaluator's judgments of my performance using student growth percentiles was similar to his/her judgments of my performance based on classroom observations" was 2.66. Both of these means were above 2.5, indicating a general level of agreement among participants. Both scores, however, are questionable because of the lower number of participants who were eligible to answer it (n = 63) and the relatively large standard deviations for each statement (1.560 and 1.591 respectively). Similar to the results of the regression analysis for utility that identified previous evaluation experience as a noteworthy coefficient, these results could suggest that teachers may view the practice more positively the more they are exposed to it.

While a mean of 2.5 for an individual statement indicates a balance of agreement and disagreement, a mean of 2.0 indicates a negative attitude. With that in mind, it is

important to note that 5 of the 11 accuracy statements had means below 2.0. The lowest mean in the domain and in the entire study was for the statement, "Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of community and cultural beliefs and practices on student achievement" (mean = 1.83). The research literature has also noted that the influence of community and culture can be difficult to account for especially as researchers and policy-makers attempt to account for students from impoverished backgrounds (Baker, et al., 2010; Callier, 2010; Fenstermacher & Richardson, 2005; Rivkin et al., 2005). The open-ended responses on the survey also noted that districts that are not funded as well as others make it difficult to compare results the way student growth percentiles do. The other 4 low scoring means centered on the accounting for school resources (mean = 1.94), school scheduling (mean = 1.96), student characteristics (mean = 1.96), and support personnel (mean = 1.99). In regards to school resources and support personnel, the open-ended responses centered on how equally these supports were available across classrooms, schools, and divisions, a concern also expressed in the literature (Baker, et al., 2010; Darling-Hammond, 2009; Greenwald, Hedges, & Laine, 1996). Also represented in the open-ended responses was how student grouping is rarely randomized and the academic propensity of students can have a large impact, issues brought to light by authors like Kane and Steiger (2008).

As was the case with the previous two research questions, the regression analysis identified the coefficients for experience ($p = .017$) and eligibility to receive additional pay ($p = .024$) as statistically significant. Teachers with 6+ years of experience rated the use of student growth percentiles 2.150 points lower in terms of accuracy as opposed to those teachers with 5 or fewer years of experience. Again, while teachers who reported

being eligible for additional pay rated it 4.281 points higher, the same questions arise because of the small number of participants identifying themselves as eligible (n = 17) and how the trend is reversed for those identifying themselves as being part of the Pay for Performance program (B = 2.266).

Conclusions

This findings of this study resulted in a number of conclusions regarding the uses of student growth percentiles in teacher evaluation.

1. Teachers in this study had an overall negative attitude toward the uses of student growth percentiles in teacher evaluation. This is evidenced by the relatively low composite scores calculated for each of the three domains of utility, propriety, and accuracy.
2. Teachers in this study were most concerned about the accuracy of using student growth percentiles in teacher evaluation. The composite score for accuracy was closest to being a full standard deviation from the identified score for neutrality, and the means for the individual statements under the accuracy umbrella were also the lowest in the study. Teachers expressed specific concerns about the influences of outside factors including community and cultural influences, school resources, scheduling, student characteristics, and support personnel on student achievement and, by extension, student growth percentiles.
3. The more experienced a teacher was, the more likely he/she was to rate the practice of using student growth percentiles negatively. This held true in all

three domains and was evidenced in not only the regression analyses but also the mean composite scores disaggregated by years of teaching.

4. It remains unknown how financial incentives play a role in teachers' attitudes toward the use of student growth percentiles due to mixed, and seemingly contradictory, findings. While teachers who reported being eligible to receive additional pay tended to rate the practice higher than their counterparts, teachers who specifically reported being a part of the Virginia Pay for Performance program, tended to rate the practice lower.
5. Teachers who had been previously evaluated with student growth percentiles rated the practice higher than those who had not been or did not know if they had been.
6. An unexpected observation was that teachers may not adequately understand what student growth percentiles are and what the timelines are for using them in a teacher's evaluation. Many of the open-ended responses reflected this lack of understanding. In addition, teachers had a relatively positive attitude in terms of student growth percentiles making mathematical sense but continued to have concerns over the accuracy of the calculations. Teachers also gave a relatively high score toward student growth percentiles allowing for timely and functional reporting despite reports that SOL scores and related reports were not made available at the end of a school year.
7. Teachers were specifically concerned about the effect that using student growth percentiles would have on the levels of collegiality among teachers in a school. This was evident in both the closed and open-ended responses.

8. Teachers appear to question the impact that using student growth percentiles has on instruction. While the score for promoting a sound education was relatively high when compared to other items, teachers were not as positive in believing that the standardized tests measure what a student should learn. In addition, comments in the open-ended sections of the survey expressed concern over "teaching to the test" and focusing on lower level thinking skills.
9. There is a large discrepancy in teachers' perceptions of how well qualified their supervisors are in terms of using student growth percentiles. While "evaluator qualifications" was one of the highest scoring items on the survey, it also had one of the largest standard deviations. In addition, various comments made by teachers alluded to their administrators not understanding student growth percentiles or not having the necessary experience to interpret them in context.

Recommendations for Practice

This study has resulted in a number of observations regarding the use of student growth percentiles in a teacher's evaluation, leading to the recommendations for action list below.

1. The foremost concern among participants in this study concerned the accuracy when using student growth percentiles in a teacher's evaluation. Teachers in general felt that there were too many factors that were out of their control and that could not be accounted for when using student growth percentiles. The most obvious need, therefore, is to study whether student growth percentiles are, indeed, an accurate measure of student academic gains. Considering

teachers expressed doubt that outside influences were taken into account, it may be worthwhile to explore a value-added approach as opposed to student growth percentiles. This, of course, comes with its own set of potential pitfalls and should be considered with caution. What may be an even more valuable conversation among policy makers is if SOL tests can be used as measures of growth in the first place. Part of this conversation would need to center on whether the tests are adequately measuring the growth of student achievement. In addition, considering that student academic progress accounts for 40 per cent of a teacher's evaluation in Virginia, it would be prudent to reexamine if this weight should be lessened considering the concerns about accuracy until empirical evidence demonstrates their predictive validity.

2. There appears to be considerable confusion over what student growth percentiles are exactly. This is especially important because a teacher evaluation system is often only as valuable as the trust that people have in it. The VDOE has already created some training modules for teachers and administrators but the real work probably will be done in face-to-face meetings with school personnel. For districts deciding to use student growth percentiles, it is vital for teachers to receive clear, hands-on training on what student growth percentiles are. This training would need to be particularly tailored for experienced teachers as one finding from the study is that the more experienced a teacher is, the more he/she does not trust the use of student growth percentiles.

3. If student growth percentiles are deemed sufficiently accurate and school personnel clearly understand them, then it is also important to find a way to apply them across all grade levels and content areas. There is a tone of resentment among some teachers that only select teachers are evaluated using student growth percentiles and in a fair system, everyone would be evaluated using similar methods.
4. If student growth percentiles are used in a teacher's evaluation, it should be one of various data points to be used. One of the expressed concerns is that SOL scores and student growth percentiles are the result of one test on one day in a year. Other data points that could be used include work samples, student portfolios, performance tasks, teacher goal setting, and division created assessments.
5. If student growth percentiles are to truly be a valuable method for evaluating teachers, they must become available sooner. Indications are that teachers and administrators do not have the data in a timely enough manner to use them thoroughly and accurately in a summative evaluation for a given school year.
6. If all of the above are resolved, school administrators must still stay cognizant of the effect that using student growth percentiles in teacher evaluation has on a school's culture. No matter what the measure is, if it creates a sense of competition, resentment, and fear, the work that teachers need to complete will not be possible. It is becoming increasingly important that teachers collaborate with each other and the decision to using a certain evaluation tool should not interfere with this need.

Recommendations for Further Research

This study helps to fill a gap in the research on using student achievement in teacher evaluation; however, it also has highlighted a number of other issues that need further research. Below are specific recommendations for further study.

1. As this study focused only on how schools in the Commonwealth of Virginia use student growth percentiles in a teacher's evaluation, it would prove helpful to perform a similar study in other states. First, this could help inform the conversation about the standardized tests used to produce student growth percentiles. Second, it could also elucidate how different states may have different standards for student academic performance. Above all, if similar results are found, it could imply that teachers' attitudes toward the use of student growth percentiles transcend specific tests and academic standards.
2. One area that became increasingly unclear during the study was that it was not known how student growth percentiles were actually used in an evaluation. For example, it was not clear if administrators had a specific numerical target for teachers or not. Or did teachers use them as part of a goal setting process? How student growth percentiles are used and analyzed could determine a teacher's attitude toward the practice.
3. Considering that student growth percentiles are just one way to measure student academic progress in Virginia, it would helpful to research teachers' attitude and beliefs about other methods. This research would help clarify if teachers' beliefs are generally negative about the general philosophy of using student academic progress in an evaluation or if it is specifically tied to student growth percentiles.

4. While the sample for this study is a rich one, a similar study with a larger sample size would be valuable. First, a study that is larger in scale could either confirm or refute the findings of this study. Second, it could explore the impact of independent variables such as eligibility for additional pay, as this study had too small of a population to produce trustworthy conclusions regarding the variable.
5. The target population for this study included teachers who had participated in the Virginia Pay for Performance program because they were the only publically known teachers evaluated using student growth percentiles. Based on the responses provided, however, only a small number of teachers indicated knowing that their school was participating. This could be the result of those in the program simply not participating in the survey or it could be that they did not know they were part of the program. A study that is able to confidently report on these teachers' perceptions would prove beneficial.
6. The open-ended response portion of the survey used in this study allowed for a richer understanding of teacher's beliefs and perceptions. Some of the comments, however, revealed that teachers have varying degrees of understanding of student growth percentiles. An in-depth, qualitative analysis of teachers' understandings of the practice and the reasons for it could help further add to the research and aid school administrators in developing appropriate professional development on using student growth percentiles in teacher evaluation.
7. At the heart of teaching is the actual instructional practice that occurs on a day-to-day basis. The research and observations in this study indicate that using student growth percentiles could have a significant impact on the quality of instruction

that students receive with particular concerns expressed over the practice of "teaching to the test" and focusing on lower level thinking skills. A study that analyzes this relationship and compares it to other methods of measuring student academic growth and their relationships would prove most revealing.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics*, *25*(1)*,* 95-135.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65-75.

American Federation of Teachers. (2010). *A continuous improvement model for teacher development and evaluation*. Washington D.C.: Author. Retrieved from http://www.aft.org/pdfs/teachers/improvemodelwhitepaper011210.pdf

Anderman, E. M., Anderman, L., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, *45*(2), 123-137.

Babbie, E. (2001). *The practice of social research* (9th ed.). Belmont, CA: Wadsworth.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers*. EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Ballard, K., & Bates, A. (2008). Making a connection between student achievement, teacher accountability, and quality classroom instruction. *Qualitative Report*, *13*(4)*,* 560-580.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-66.

Bennett, W. (January 11, 2012). The lasting impact of good teachers. *CNN Opinion*. Retrieved from http://www.cnn.com/2012/01/11/opinion/bennett-good-teachers

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*. *56*, 205-213.

Betenbenner, D., & Linn, R. (December, 2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, Princeton, NJ.

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *Elementary School Journal*, *106*(1), 3-20.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting the value out of value-added: Report of a workshop*. Washington DC: The National Academies Press.

Briggs, A., & Coleman, M. (Eds.). (2007). *Research methods in educational leadership and management.* Los Angeles, CA: Sage Publications.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36*(5), 616-637.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*(10), 1069-1077.

Callier, J. (2010). Paying teachers according to student achievement: Questions regarding pay-for-performance models in public education. *Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(2), 58-61.

Cantrell, S., & Kane, T. (2013). *Ensuring fair and reliable measures of effective teaching.* Policy Brief. MET Project. Seattle, WA: The Bill and Melinda Gates Foundation

Cantrell, S., & Kane, T. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Policy Brief. MET Project. Seattle, WA: The Bill and Melinda Gates Foundation

Carey, K., & Manwaring, R. (2011). *Growth models and accountability: A recipe for remaking ESEA*. Washington DC: Education Sector.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Education Evaluation and Policy Analysis, 24*(4), 305-331.

Center for Public Education. (2007). *Measuring student growth: At a glance*. Alexandria, VA: Hull, J. Retrieved from: http://www.centerforpubliceducation.org/Main-Menu/Policies/Measuring-student-growth-At-a-glance.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education.* London: Routledge. doi:10.4324/9780203224342.

Cooper, B. S., Ehrensal, P. L., & Bromme, M. (2005). School-level politics and professional development: Traps in evaluating the quality of practicing teachers. *Educational Policy*, *19*(1), 112-125.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and*

practice. Executive Summary. Education Policy for Action Series. Providence, RI: Annenberg Institute for School Reform at Brown University.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*, 464-494. doi:10.1086/318641

Cowart, B., & Myton D. (1997). The Oregon teacher work sample methodology: Rationale and background. In J. Milman (Ed.), *Grading teachers, grading schools.* (pp. 11-14) Thousand Oaks, CA: Corwin Press.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: ASCD.

Darling-Hammond, L. (2009). Recognizing and enhancing teacher effectiveness. *The International Journal of Educational and Psychological Assessment*, *3*, 1-24.

Darling-Hammond, L., & And, O. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, *53*(3), 285-328.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education*, *104*(2), 289-319.

Derrington, L. (2011). Changes in teacher evaluation: Implications for the principal's work. *Delta Kappa Kamma Bulletin*, *77*(3), 51.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of Internet-based surveys: An experimental study. *Marketing Letters, 15*(1), 21-36.

Dillman, D., & Bowker, D. (2001). The web questionnaire challenge to survey methodologists. In U. Reips & M. Bosnjak, (Eds.), *Dimensions of Internet science.* Lengerich, Germany: Pabst Science.

Dillman, D., Tortora, R., Conradt, J., & Bowker, D. (1998). *Influence of plain vs. fancy design on response rates for web surveys*. Paper presented at the Joint Statistical Meetings, Dallas, TX.

Dillman, D., Tortora, R., & Bowker, D. (2001). *Principles for constructing web surveys.* Working paper.

Dillman, D. (1999). *Mail and Internet surveys: The tailored design method.* New York: J. Wiley.

Dillman, D. (2000*). Mail and Internet surveys: The tailored design method (2nd ed.).* New York: J. Wiley.

Ding, C., & Sherman, H. (2006). Teaching effectiveness and student achievement: Examining the relationship. *Educational Research Quarterly*, *29*(4), 40-51.

Donaldson, M. (2012). *Teachers' perspectives on education reform.* Washington DC: Center for American Progress.

Duke, D. (2008). Diagnosing school decline. *Phi Delta Kappan*, *89*(9), 667-671.

Duncan, A. (2012). Virginia waiver approval letter. Retrieved from http://www.doe.virginia.gov/federal_programs/esea/flexibility/secretarys_approval_letter.pdf

Evans, J., & Mathur, A. (2005). The value of online surveys. *Internet Research, 15*(2), 195-219. doi:10.1108/10662240510590360

Felner, R. D., Bolton, N., Seitsinger, A. M., Brand, S., & Burns, A. (2008). Creating a statewide educational data system for accountability and improvement: A comprehensive information and assessment system for making evidence-based change at school, district, and policy levels. *Psychology in the Schools*, *45*(3), 235-256.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186-213.

Fink, A. (2002). *How to manage, analyze, and interpret survey data* (2nd ed.). Thousand Oaks, CA: Sage.

Fogelman, K., & Comber, C. (2007). Surveys and sampling. In Briggs, A. & Coleman, M., eds. *Research methods in educational leadership and management*. (pp.125-141). Los Angeles, CA: Sage Publications.

Frymier, J. (1998). Accountability and student learning. *Journal of Personnel Evaluation in Education*, *12*(3), 233-35.

Gallagher, A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement?. *Peabody Journal of Education*, *79*(4), 79-107.

Goe, L., Holdheide, L., Miller, T., & National Comprehensive Center for Teacher, Q. (2011). A practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems. Washington DC: *National Comprehensive Center for Teacher Quality*.

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (October, 2005). *Policymakers' guide to student growth models for school*

accountability: How do accountability models differ? Washington D.C.: Council for Chief State School Officers.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 361-96.

Groves, R., Dillman, D., Eltinge, J., Little, R., Biemer, P., & Groves, R. (2005). Survey methodology. *Technometrics, 47*(2), 246-246.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey methodology,* 2nd ed. Hoboken, NJ: John Wiley & Sons.

Gullickson, Arlen R. (2009). *The Personnel Evaluation Standards: How to assess systems for evaluating educators.* (2nd ed.) Thousand Oaks, CA: Corwinn Press.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis*, *8*(1), 45-60.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, *33*(3), 329-351.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, *61*(2), 280-288.

Hanushek, E. A. (2011). Valuing teachers: How much is a good teacher worth? *Education Next*, *11*(3), 41-45.

Hanushek, E. A., & Rivkin, S. G. (2007). Pay, working conditions, and teacher quality. *Future Of Children*, *17*(1), 69-86.

Hanushek, E. A., & Rivkin, S. G (2010). *Using value-added measures of teacher quality*. Brief 9. Washington DC: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* Working Paper 30. Washington D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Harris, D., & McCaffery, D. (2010). Value-added: Assessing teachers' contributions to student achievement. In Kennedy, M. (Ed.), *Teacher assessment and the quest for teacher quality.* (pp. 251-282). San Francisco, CA: Jossey-Bass.

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Guidance for developing and selecting assessments of student growth for use in teacher evaluation systems (Extended Version)*. Los Angeles, CA: Assessment and Accountability Comprehensive Center.

Hess, F. M., & Kelly, A. P. (2007). Learning to lead: What gets taught in principal-preparation programs. *Teachers College Record*, *109*(1), 244-274.

Hinchey, P. H. (2010). *Getting teacher assessment right: What policymakers can learn from research*. Boulder, CO: National Education Policy Center.

Holdheide, L. R., Goe, L., Croft, A., & Reschly, D. J. (2010). *Challenges in evaluating special education teachers and English language learner specialists*. Research & Policy Brief. Washington D.C.: National Comprehensive Center for Teacher Quality.

Ingvarson, L., & Rowe, K. (2008). Conceptualizing and evaluating teacher quality: Substantive and methodological issues. *Australian Journal of Education*, *52*(1), 5-35.

Jacob, B., & Lefgen, L. (2006). When principals rate teachers: The best--and the worst--stand out. *Education Next*, *6*(2), 58-64.

Jacob, B., & Walsh, E. (2011). What's in a rating? *Economic of Education Review, 30*(3), 434-448.

Johnson, C., Kahle, J., & Fargo, J. (2007). Effective teaching results in increased science achievement for all students. *Journal of Science Education, 91*(3), 371-383.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research.

Kane, T.J., Wooten, A.L., Taylor, E.S., & Tyler, J. (2011). Evaluating teacher effectiveness. *Education Next*, *11*(3).

Kaplowitz, M., Hadlock, T., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly, 8*(1), 94-101.

Kennedy, M. (Ed.). (2010). *Teacher assessment and the quest for teacher quality: A handbook*. Hoboken, NJ. Jossey-Bass.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal Of Personnel Evaluation in Education*, *16*(4), 241-68.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34-70.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper 2007-03. Nashville, TN: National Center on Performance Incentives.

Konstantopoulos, S. (2009). Effects of teachers on minority and disadvantaged students' achievement in the early grades. *Elementary School Journal*, *110*(1), 92-113.

Krathwol, D. R. (1998). *Methods of educational and social science research: An integrated approach* (2nd ed.). New York: Longman.

Lai, F., Sadoulet, E., & de Janvry, A. (2011). The contributions of school quality and teacher qualifications to student performance: Evidence from a natural experiment in Beijing middle schools. *Journal of Human Resources*, *46*(1), 123-153.

Larson, P., & Chow, G. (2003). Total cost/response trade-offs in mail survey research: Impact of follow-up mailings and monetary incentives. *Industrial Marketing Management, 32*, 533-537.

Linn, R. L. (2005). Issues in the design of accountability systems. *Yearbook of The National Society For The Study Of Education*, *104*(2), 78-98.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to different

mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47-67.

Loup, K. S., & And, O. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, *10*(3), 203-26.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86*(10), 727.

Marshall, K. (2009). *Rethinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap*. Hoboken, NJ. Jossey-Bass.

Marzano, R. (2012). The two purposes of teacher evaluation. *Educational Leadership, 70*(3)*,* 14-19.

McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project.* Technical Report. TR-506-CC. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.

McConney, A., Schalock, M. D., & Schalock, H. D. (1997). Indicators of student learning in teacher evaluation. In J. H. Stronge (Ed.) *Evaluating teaching: A guide to current thinking and best practice.* (pp. 162-192) Thousand Oaks, CA: Corwin Press.

Mendro, R. L. (1998). Student achievement and school and teacher accountability. *Journal of Personnel Evaluation in Education*, *12*(3), 257-67.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Milanowski, A. (2011). Strategic measures of teacher performance. *Phi Delta Kappan*, *92*(7), 19-25.

Milanowski, A. T., & Heneman, H. H. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, *15*(3), 193-212.

Misco, T. (2008). Was that a result of my teaching? A brief exploration of value-added assessment. *Clearing House: A Journal Of Educational Strategies, Issues and Ideas*, *82*(1), 11-14.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research & Evaluation*, *12*(1), 53-74.

Munoz, M. A., & Chang, F. C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal Of Personnel Evaluation In Education*, *20*(3-4), 147-164.

National Council on Teacher Quality. (October, 2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies.* Washington DC: NCTQ.

National Education Association. (2010). Teacher assessment and evaluation: The National Education Association's framework for transforming education systems

to support effective teaching and student learning. Retrieved from http://www.nea.org/assets/docs/HE/TeachrAssmntWhtPaperTransform10_2.pdf

National Governors Association. (2011). *Preparing principals to evaluate teachers.* Washington D.C.: National Governors Association.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23), 1-24.

Nolan, J. F., & Hoover, L. A. (2004). *Teacher supervision and evaluation: Theory into practice.* New Jersey: Wiley.

O'Malley, K., Murphy, S., McLarty, K., Murphy, D., & McBride, Y. (2011). *Overview of growth models.* Upper Saddle River, NJ: Pearson.

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement*, *70*(3), 401-419.

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*. *48*(1), 163-193.

Patten, Mildred L. (2010). *Proposing empirical research: A guide to the fundamentals.* (4th ed.) Glendale, CA: Pyrczak Publishing.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin Press.

Peterson, K. (2004). Research on school teacher evaluation. *NASSP Bulletin*, *88*(639), 60-79.

Peterson, K. D., & Peterson, C. A. (2006). *Effective teacher evaluation: A guide for principals.* Thousand Oaks, CA: Corwin.

Phillips, K. R. (2010). What does "highly qualified" mean for student achievement? Evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *Elementary School Journal*, *110*(4), 464-493.

Porter, S. R. (2004). Raising response rates. *New directions for institutional research*. *2004*(121). 5-21.

Pryor, J., (2004). Conducting surveys on sensitive subjects. In S.R. Porter (Ed.) *New directions for institutional research: No. 121 Overcoming survey research problems* (pp. 39-50). San Francisco, CA: Jossey-Bass.

Ravitch, D. (2010). *The death and life of the great American school system: how testing and choice are undermining education.* New York, NY: Basic Books.

Rivkin, S., Hanushek, E., & Kain, J.F. (2005). Teacher, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2003). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

Rockoff, J., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18*(5), 687-696.

Rosenberg, S. & Silva, E. (2012). *Trending toward reform: Teachers speak on unions and the future of the profession.* Washington DC: Education Sector Reports.

Sanders, W., Saxton, A., & Horn, S. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In

Ed. J. Milman *Grading teachers, grading schools.* (pp. 137-162) Thousand Oaks, CA: Corwin Press.

Sanders, W. L., & Rivers, J. C. (1996, November). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*. Technical Report. Santa Monica, CA: RAND Corporation.

Stiggins, R. (2001). The principal's role in assessment. *NASSP Bulletin*, *85*, 13-26.

Stiggins, R. (1998). *Classroom assessment for student success*. Annapolis Junction, MD: NEA Publications.

Stiggins, R., & Duke, D. (1988). *The case for commitment to teacher growth.* Albany, NY: SUNY Press.

Stronge, J. (Ed.). (2006). *Evaluating teaching*. (2nd ed.) Thousand Oaks, CA: Corwin Press.

Stronge, J. (1997). Improving schools through teacher evaluation. In J. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice.* (pp. 1-23). Thousand Oaks, CA: Corwin Press.

Stronge, J., & Tucker, P. (2000). *Teacher evaluation and student achievement.* Annapolis Junction, MD: NEA Publications.

Stronge, J., Ward, T. J., & Grant, L. (2011). What makes good teachers good?: A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.

Stronge, J., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2007). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education*, *20*(3-4), 165-184.

Stronge, J., Ward, T. J., Tucker, P. D., Hindman, J. L., McColsky, W., & Howard, B. (2007). National board certified teachers and non-national board certified teachers: Is there a difference in teacher effectiveness and student achievement? *Journal of Personnel Evaluation in Education*, *20*(3-4), 185-210.

Stufflebeam, D. (1997). "Oregon teacher work sample methodology: Educational policy review. In Ed. J. Milman *Grading teachers, grading schools.* (pp. 219-226). Thousand Oaks, CA: Corwin Press.

Teoh, M. (2012). *Great expectations: Teachers views on elevating the teaching profession.* Boston, MA: TeachPlus.

Thum, Y., & Bryk, A. (1997). Value-added productivity indicators: The Dallas System. In J. Milman (Ed.) *Grading teachers, grading schools.* (pp. 100-108) Thousand Oaks, CA: Corwin Press.

United States Department of Education. (2010). *A blueprint for reform: The reauthorization of the elementary and secondary education act.* Washington, DC: USDOE.

United Stated Department of Education. (2011). *Obama administration sets high bar for flexibility from No Child Left Behind in order to advance equity and support reform.* Washington D.C.: USDOE.

Valli, L., Croninger, R. G., & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education*, *113*(4), 635-662.

Virginia Department of Education. (2011). 2011 board of education teacher performance standards & evaluation criteria. Retrieved from http://www.doe.virginia.gov/teaching/performance_evaluation/teacher/index.shtml.

Virginia Department of Education. (2012). Accountability and Virginia schools. Retrieved from http://www.doe.virginia.gov/statistics_reports/school_report_card/accountability_guide.shtml

Virginia Department of Education. (2011). *Guidelines for the uniform performance standards and evaluation criteria for teachers.* Richmond, VA: USDOE.

Virginia Department of Education. (2011). Governor McDonnell announces performance- pay pilot schools. Retrieved from http://www.doe.virginia.gov/news/news_releases/2011/july21_gov.shtml.

Virginia Department of Education. (2011). Student growth percentiles. Retrieved from http://www.doe.virginia.gov/testing/scoring/student_growth_percentiles/index.shtml.

Virginia Education Association. (2011). Teacher evaluation guidelines raise question of time, resources. Retrieved from http://www.veanea.org/home/1210.htm.

Virginia School Laws. (2012). *Code of Virginia.* § 22.1-253.13:5. Charlottesville, VA: Michie.

Virginia School Laws. (2012). *Code of Virginia.* § 22.1-295. Charlottesville, VA: Michie.

Webb, N. L. (2002, December). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Webster, W. & Mendro, R. (1997). The Dallas value-added accountability system. In Ed. J. Milman *Grading teachers, grading schools.* (pp. 81-98) Thousand Oaks, CA: Corwin Press.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Executive Summary. Second Edition. New York: New Teacher Project.

Weisberg, H.F., Krosnick, J., & Bowen. B. (1996). *An introduction to survey research, polling, and data analysis*. New York: Sage Publications, Inc.

Whitcomb, M., & Porter, S. (2004). E-Mail contacts: A test of complex graphical designs in survey research. *Social Science and Computer Review, 22*(3), 370-376. doi:10.1177/0894439304263590

White, B. (April, 2004). *The relationship between teacher evaluation scores and student achievement: Evidence from Coventry, RI*. Paper presented at the American Education Research Association national conference, San Diego, CA.

Wright, S., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57-67.

Yeh, S. S., & Ritter, J. (2009). The cost-effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance*, *34*(4), 426-451.

Zimmerman, S., & Deckert-Pelton, M. (2003). Evaluating the evaluators: Teachers' perceptions of the principal's role in professional evaluation. *NASSP Bulletin*, *87*(636), 28-37.

Appendix A: Tables

Table 22

Rank Order of Survey Statements by Mean

Factor	N	Min	Max	Mean	Std Dev.
Allows for adequate accounting of community and cultural influences.	145	1	4	1.83	.890
Fosters sense of collegiality	148	1	4	1.89	.907
Allows for adequate accounting of school resources.	144	1	4	1.94	.868
Allows for adequate accounting of school scheduling.	146	1	4	1.96	.917
Allows for adequate accounting of student characteristics.	144	1	4	1.96	.982
Reflects teacher's strengths and weaknesses.	148	1	4	1.97	.903
Allows for adequate accounting of influence of support personnel.	145	1	4	1.99	.861
Standardized tests adequately measure what students should learn.	145	1	4	2.10	.964
Reliability.	145	1	4	2.12	.887
Aids in retention of good teachers.	149	1	4	2.17	.942
Allows for adequate accounting of influence of previous teachers.	148	1	4	2.18	.864
Promotes education of all students.	150	1	4	2.21	.963
Promotes sense of professionalism.	146	1	4	2.21	.946
Provides clear expectations.	141	1	4	2.23	.981

Student growth percentiles make mathematical sense.	145	1	4	2.24	.880
Promotes sound education.	150	1	4	2.30	.915
Reinforces good instructional practice.	148	1	4	2.36	.934
Timely and functional reporting.	64	1	4	2.40	1.303
Evaluation judgments with student growth percentiles were similar to classroom observations.	63	1	4	2.66	1.591
Evaluator qualifications.	143	1	4	2.73	1.048
Summative score with student growth percentiles was similar to previous years.	63	1	4	2.85	1.560

Table 23

Composite Scores for Propriety, Utility, and Accuracy

Domain	N	Min	Max	Mean	Std.Dev.
Propriety	149	3	12	6.50	2.511
Utility	149	5	26	14.26	5.094
Accuracy	149	0	39	21.39	7.382

Table 24

Composite Scores Disaggregated by Experience

Years	Domain	N	Min	Max	Mean	Std.Dev.
0-5	Propriety	32	3	12	7.09	2.607
	Utility	32	5	26	16.47	4.813
	Accuracy	32	8	39	23.44	7.457
6-20	Propriety	81	3	12	6.51	2.491
	Utility	81	5	24	14.25	5.152
	Accuracy	81	0	35	21.57	7.500
20+	Propriety	36	3	12	5.97	2.420
	Utility	36	6	24	12.33	4.504
	Accuracy	36	10	35	19.17	6.605

Table 25

Composite Scores Disaggregated by Professional Affiliation

Affiliation	Domain	N	Min	Max	Mean	Std.Dev.
Yes	Propriety	66	3	12	6.44	2.678
	Utility	66	5	26	14.23	5.549
	Accuracy	66	0	39	21.39	7.840
No	Propriety	83	3	12	6.55	2.385
	Utility	83	5	24	14.29	4.736
	Accuracy	83	8	35	21.39	7.045

Table 26

Composite Scores Disaggregated by Previous Evaluation Experience

Previously Evaluated w/ SGP	Domain	N	Min	Max	Mean	Std.Dev
Yes	Propriety	53	3	12	7.04	2.519
	Utility	53	6	24	15.96	4.747
	Accuracy	53	9	35	23.08	7.022
No/Don't Know/Didn't Teach	Propriety	96	3	12	6.21	2.471
	Utility	96	5	26	13.32	5.059
	Accuracy	96	0	39	20.46	7.447

Table 27

Composite Scores Disaggregated by Eligibility to Receive Additional Pay

Eligibility	Domain	N	Min	Max	Mean	Std.Dev
Yes	Propriety	19	3	12	8.21	2.417
	Utility	19	8	24	18.00	5.066
	Accuracy	19	0	35	25.42	8.553
No/Don't Know/Didn't Teach	Propriety	130	3	12	6.25	2.435
	Utility	130	5	26	13.72	4.882
	Accuracy	130	6	39	20.80	7.041

Table 28

Composite Scores Disaggregated by Participation in Pay for Performance Program

Participation	Domain	N	Min	Max	Mean	Std.Dev
Yes	Propriety	17	3	11	6.35	2.849

	Utility	17	6	24	13.06	5.728
	Accuracy	17	9	33	20.06	6.619
No/Don't Know/Didn't Teach	Propriety	132	3	12	6.52	2.476
	Utility	132	5	26	14.42	5.010
	Accuracy	132	0	39	21.56	7.481

Table 29

Cronbach's Alpha Value for Composite Scores

Domain	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
Propriety	.878	.878	3
Utility	.825	.842	7
Accuracy	.836	.871	11

Appendix B: Survey Instrument

This survey is designed to assess your perceptions of the use of student growth percentiles in teacher evaluation. In this section, the survey asks you about you and your experience as a teacher:

1) What grade level did you primarily teach in 2011-2012?
   1. 4th Grade
   2. 5th Grade
   3. 6th Grade
   4. 7th Grade
   5. 8th Grade
   6. Did not teach in 2011-2012

2) Which of the following subject(s) did you teach in 2011-2012?
   1. Math
   2. Reading
   3. Math and Reading
   4. Did not teach in 2011-2012

3) How many years have you been teaching?
   1. Less than 3 years
   2. 3-5 years
   3. 6-10 years
   4. 11-20 years
   5. 20+ years

4) Are you a member of the NEA (National Education Association) or AFT (American Federation of Teachers)?
   1. Yes
   2. No

5) Did your evaluator use student growth percentiles as a source of information when completing your summative evaluation for the 2011-2012 school year?
   1. Yes
   2. No
   3. Don't know
   4. Did not teach in 2011-2012

6) Were you eligible to receive additional compensation based on your 2011-2012 summative evaluation?

1. Yes
2. No
3. Don't know
4. Did not teach in 2011-2012

7) Is your school participating in the Virginia Pay for Performance initiative?

1. Yes
2. No
3. Don't know

The Commonwealth of Virginia will soon require that 40% of all teacher evaluations be based on "student academic progress." One of the ways schools can measures "student academic progress" is through the use of student growth percentiles that are provided by the Virginia Department of Education (VDOE). Important specifics about student growth percentiles include:

- They are provided for students in math and reading in Grades 4-9.
- They show how a student's SOL score compares to other students who have scored similarly the previous year. Students receive a score ranging from 1 to 99. A low score indicates the student performed poorly as compared to other students who scored similarly the previous year, whereas a high score indicates the student performed well in comparison.
- Student growth percentiles do not account for whether a student passed – just how he/she compares to other students with similar past SOL scores.
- Student growth percentiles are not calculated for students who score an advanced pass score (500 or above).

This survey is designed to assess your perceptions of the use of student growth percentiles in teacher evaluation.

In this section, please rate to what degree you either agree or disagree with each of the following statements. At the end of each of the section, there is an opportunity for you to expand on your perceptions and opinions with an open-ended question.

1) Using student growth percentiles in teacher evaluation promotes a sound education for students.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

2) Using student growth percentiles in teacher evaluation promotes the education of all students.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

3) Using student growth percentiles in teacher evaluation allows for the strengths and weaknesses of a teacher to be shown in the evaluation.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

Using student growth percentiles in a teacher's evaluation should be done in a legal and ethical manner, with the welfare of the evaluatee in mind. Please provide any additional comments you may have regarding this stance.

The Commonwealth of Virginia will soon require that 40% of all teacher evaluations be based on "student academic progress." One of the ways schools can measures "student academic progress" is through the use of student growth percentiles that are provided by the Virginia Department of Education (VDOE). Important specifics about student growth percentiles include:

- They are provided for students in math and reading in Grades 4-9.
- They show how a student's SOL score compares to other students who have scored similarly the previous year. Students receive a score ranging from 1 to 99. A low score indicates the student performed poorly as compared to other students who scored similarly the previous year, whereas a high score indicates the student performed well in comparison.
- Student growth percentiles do not account for whether a student passed – just how he/she compares to other students with similar past SOL scores.
- Student growth percentiles are not calculated for students who score an advanced pass score (500 or above).

In this section, please rate to what degree you either agree or disagree with each of the following statements. At the end of each of the section, there is an opportunity for you to expand on your perceptions and opinions with an open-ended question.

4) Using student growth percentiles in teacher evaluation will aid in the retention of good teachers.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

5) Using student growth percentiles in teacher evaluation reinforces good instructional practice.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

6) Using student growth percentiles in teacher evaluation promotes a sense of professionalism in a school.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

7) Using student growth percentiles in teacher evaluation helps foster a sense collegiality among staff members at a school.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

8) It was clear what was expected of a teacher in terms of student achievement as it is reflected in the student growth percentiles.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

9) Student growth percentile data was reported in a timely and functional way to be used in my summative evaluation. (NOTE: Mark Not Applicable if student growth percentiles were not used in your evaluation in 2011-2012 or if you did not teach in 2011-2012.)

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree
5. Not Applicable

10) My evaluator is qualified and capable of interpreting student growth percentiles data in my evaluation.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

Using student growth percentiles in a teacher's evaluation should allow the evaluation to be informative, timely, and influential. Please provide any additional comments you may have regarding this stance.

The Commonwealth of Virginia will soon require that 40% of all teacher evaluations be based on "student academic progress." One of the ways schools can measures "student academic progress" is through the use of student growth percentiles that are provided by the Virginia Department of Education (VDOE). Important specifics about student growth percentiles include:

- They are provided for students in math and reading in Grades 4-9.
- They show how a student's SOL score compares to other students who have scored similarly the previous year. Students receive a score ranging from 1 to 99. A low score indicates the student performed poorly as compared to other students who scored similarly the previous year, whereas a high score indicates the student performed well in comparison.
- Student growth percentiles do not account for whether a student passed – just how he/she compares to other students with similar past SOL scores.
- Student growth percentiles are not calculated for students who score an advanced pass score (500 or above).In this section, please rate to what degree you either agree or disagree with each of the following statements.

At the end of the section, there is an opportunity for you to expand on your perceptions and opinions with an open-ended question.

11) My evaluator's judgments of my performance using student growth percentiles was similar to his/her judgments of my performance based on classroom observations. (NOTE: Please mark Not Applicable if you were not evaluated using student growth percentiles in 2011-2012 or did not teach in 2011-2012.)

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

5. Not Applicable

12) My summative (final) rating in 2011-2012 was similar to summative ratings I have received in the past. (NOTE: Please select Not Applicable if you did not teach in 2011-2012.)
1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree
5. Not Applicable

13) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of previous years' teachers on students' current year's achievement.
1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

14) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of support personnel (e.g., reading specialists, special education teachers, etc.) on students' achievement.
1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

15) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of school scheduling (e.g., class size, class diversity, etc.) on student achievement.
1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

16) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of school resources (e.g., curricula, textbooks, furniture, etc.) on student achievement.
1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

17) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of student characteristics, especially ones associated with "at-risk" students, on student achievement.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

18) Using student growth percentiles in teacher evaluation allows for the adequate accounting of the influence of community and cultural beliefs and practices on student achievement.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

19) The standardized tests used to determine a student's student growth percentiles adequately measure what a student should learn.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

20) Student growth percentiles are a reliable tool to use in teacher evaluation (note: "Reliability" refers to how consistent the results are, not whether the results are correct or not.)

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

21) The way student growth percentiles are determined makes mathematical sense.

1. Strongly Agree
2. Somewhat Agree
3. Somewhat Disagree
4. Strongly Disagree

Using student growth percentiles in a teacher's evaluation should be based on the goal of providing sound information for the purpose of making sound judgments. Please provide any additional comments you may have regarding this stance.

Appendix C: Informed Consent Agreement

Informed Consent Agreement

Please read this consent agreement carefully before you decide to participate in the study.

Purpose of the research study: The purpose of the study is to assess teachers' perceptions of the use of student growth percentiles in teacher evaluation.

What you will do in the study: In this study you will be asked to complete a survey that measures your degree of agreement with a series of statements regarding the use of student growth percentiles in teacher evaluation. You will also be asked to respond to open-ended questions to expand on your perceptions. You can skip any question that makes you feel uncomfortable and cans top the survey at any time.

Time required: The study will require about 10-15 minutes of your time.

Risks: There are no anticipated risks in the participation of this study.

Benefits: There are no direct benefits to you for participating in this research study. The study may help us understand how to improve the practice of including student learning measures in teacher evaluation and the professional development associated with it. The study may also help in creating and revising policies related to using student learning measures in teacher evaluation.

Confidentiality: The information that you give in the study will be handled confidentially.

Voluntary participation: Your participation in the study is completely voluntary.

Right to withdraw from the study: You have the right to withdraw from the study at any time without penalty.

How to withdraw from the study: If you want to withdraw from the study, exit out of the link connected to the survey. There is no penalty for withdrawing.

Payment: You will receive no payment for participating in the study.

If you have questions about the study, contact:
Michael C. Irani
Department of Leadership, Foundations, and Policy, Curry School of Education
University of Virginia, Charlottesville, VA 22903.

Telephone: (434)466-6598
Email address: mci5m@virginia.edu

Dr. Pamela Tucker, Faculty Advisor
Department of Leadership, Foundations, and Policy, Curry School of Education
Olsson Room 226B
University of Virginia, Charlottesville, VA 22903.
Telephone: (434)924-7846
Email address: pdtucker@virginia.edu

If you have questions about your rights in the study, contact:
Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences
One Morton Dr Suite 500
University of Virginia, P.O. Box 800392
Charlottesville, VA 22908-0392
Telephone: (434) 924-5999
Email: irbsbshelp@virginia.edu
Website: www.virginia.edu/vpr/irb/sbs

Agreement:
I agree to participate in the research study described above.
Signature: _____ Date: _____
You will receive a copy of this form for your records.

Appendix D: Introductory Letter to Participants

«Salutation» «FirstName» «LastName»
«Address1»
«Address2»
«Address3»
«Address4»

April 22, 2013

Dear «Salutation» «FirstName» «LastName»,

I am writing to enlist your help in providing Virginia policy makers and school administrators a vital perspective about teacher evaluation.

As you probably know, 40% of a teacher's evaluation is soon to be based on "student academic progress." One way schools can measure "student academic progress" is through the use of "student growth percentiles" that are provided by the Virginia Department of Education.

In the next 2-4 days you will be receiving an e-mail with a link to a survey asking you to share your opinions about this approach to teacher evaluation. All responses will be kept anonymous. Your participation in this survey will not only provide Virginia policy makers and school administrators a teacher's perspective of the new teacher evaluation system, but it will also contribute to the related research in the field of teacher evaluation.

Thank you for taking part in this survey that will help inform future decisions regarding teacher evaluation in Virginia.

Sincerely,

Michael C. Irani
Doctoral Student
Department of Leadership, Foundations, and Policy
mci5m@virginia.edu

Appendix E: Initial E-mail Contact

Dear _____,

You have been selected to take part in a survey that will provide Virginia policy makers and school administrators a vital perspective about teacher evaluation. As a teacher in the Commonwealth, it is critical that your voice be heard regarding this topic.

Please use the link below to access the online survey.
<<Link>>

All responses will remain confidential. Your participation in the study will not only provide Virginia policy makers and school administrators a teacher's perspective of the new teacher evaluation policies, but it will also contribute to the related research in the field of teacher evaluation.

For further information, please complete the contact form at the link below.
<<link>>

Thank you,
Michael C. Irani

Appendix F: Subsequent E-mail Contacts

Dear _____,

This email is a reminder that you have been selected to participate in a doctoral research study assessing teachers' perceptions of the use of student growth percentiles in teacher evaluation. It is critical that your voice be heard about this very important issue.

Please use the link below to access the online survey.
<<Link>>

All responses will remain confidential. Your participation in the study will not only provide Virginia policy makers and school administrators a teacher's perspective of the new teacher evaluation policies, but it will also contribute to the related research in the field of teacher evaluation.

For further information, please complete the contact form at the link below.
<<link>>

Thank you,
Michael C. Irani

Appendix G: Approval

UNIVERSITY of VIRGINIA
OFFICE OF THE VICE PRESIDENT FOR RESEARCH
INSTITUTIONAL REVIEW BOARD FOR THE SOCIAL AND BEHAVIORAL SCIENCES

In reply, please refer to: Project # 2013-0140-00

April 22, 2013

Michael Irani and Pamela Tucker
Leadership, Foundations & Policy
PO Box 400273

Dear Michael Irani and Pamela Tucker:

Thank you for submitting your project entitled: "Teachers' Perception of the Use of Student Learning Measures in Teacher Evaluation: An Examination of the Use of Student Growth Percentiles in Virginia" for review by the Institutional Review Board for the Social & Behavioral Sciences. The Board reviewed your Protocol on April 23, 2013.

The first action that the Board takes with a new project is to decide whether the project is exempt from a more detailed review by the Board because the project may fall into one of the categories of research described as "exempt" in the Code of Federal Regulations. Since the Board, and not individual researchers, is authorized to classify a project as exempt, we requested that you submit the materials describing your project so that we could make this initial decision.

As a result of this request, we have reviewed your project and classified it as exempt from further review by the Board for a period of four years. This means that you may conduct the study as planned and you are not required to submit requests for continuation until the end of the fourth year.

This project # 2013-0140-00 has been exempted for the period April 22, 2013 to April 21, 2017. If the study continues beyond the approval period, you will need to submit a continuation request to the Board. If you make changes in the study, you will need to notify the Board of the changes.

Sincerely,

Tonya R. Moon, Ph.D.
Chair, Institutional Review Board for the Social and Behavioral Sciences

One Morton Drive, Suite 500 • Charlottesville, VA 22903
P.O. Box 800392 • Charlottesville, VA 22908-0392
Phone: 434-924-5999 • Fax: 434-924-1992
www.virginia.edu/vpr/irb/sbs.html