

Houston, We Have a Problem: Studying the SAS Education Value-Added
Assessment System (EVAAS) from Teachers' Perspectives in the Houston

Independent School District (HISD)

by

Clarín Collins

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2012 by the
Graduate Supervisory Committee:

Audrey Amrein-Beardsley, Chair
Gustavo E. Fischman
David C. Berliner

ARIZONA STATE UNIVERSITY

December 2012

ABSTRACT

This study examined the intended and unintended consequences associated with the Education Value-Added Assessment System (EVAAS) as perceived and experienced by teachers in the Houston Independent School District (HISD). To evaluate teacher effectiveness, HISD is using EVAAS for high-stakes consequences more than any other district or state in the country. A large-scale electronic survey was used to investigate the model's reliability and validity; to determine whether teachers used the EVAAS data in formative ways as intended; to gather teachers' opinions on EVAAS's claimed benefits and statements; and to understand the unintended consequences that occurred as a result of EVAAS use in HISD. Mixed methods data collection and analyses were used to present the findings in user-friendly ways, particularly when using the words and experiences of the teachers themselves.

Results revealed that the reliability of the EVAAS model produced split and inconsistent results among teacher participants, and teachers indicated that students biased the EVAAS results. The majority of teachers did not report similar EVAAS and principal observation scores, reducing the criterion-related validity of both measures of teacher quality. Teachers revealed discrepancies in the distribution of EVAAS reports, the awareness of trainings offered, and among principals' understanding of EVAAS across the district. This resulted in an underwhelming number of teachers who reportedly used EVAAS data for formative purposes.

Teachers disagreed with EVAAS marketing claims, implying the majority did not believe EVAAS worked as intended and promoted. Additionally, many unintended consequences associated with the high-stakes use of EVAAS emerged through teachers' responses, which revealed among others that teachers felt heightened pressure and competition, which reduced morale and collaboration, and encouraged cheating or teaching to the test in attempt to raise EVAAS scores.

This study is one of the first to investigate how the EVAAS model works in practice and provides a glimpse of whether value-added models might produce desired outcomes and encourage best teacher practices. This is information of which policymakers, researchers, and districts should be aware and consider when implementing the EVAAS, or any value-added model for teacher evaluation, as many of the reported issues are not specific to the EVAAS model.

DEDICATION

This dissertation is dedicated to my Grandmother, Beatrice Veronica Mahoney Collins Zoll Welch, a retired, devoted teacher for 28 years in Allamakee County, Iowa. As a young woman she developed a passion for education and teaching, and even left home to pursue and earn her high school diploma. She was the sole teacher in a one-room country schoolhouse for seven years, and spent the last 21 years of her career teaching in a public elementary school. She is proud to have been a teacher who touched the lives of hundreds of students with her intelligence, humor, and disciplined manner.

Though she was widowed while raising five young children, she supported the family on her teaching salary, and earned her college degree by attending classes on nights, weekends and during the summer. Beatrice instilled in each of her children the value and utility that earning an education can provide. Watching her children and grandchildren graduate from high school and college has brought great pride in her life.

Hearing my grandmother's voice fill with enthusiasm when I told her I was going to graduate school to study education is something I hold dear. And while her memory may be escaping her, I know in her heart she is proud of my academic accomplishment, and proud that the experiences and voices of teachers will be heard through this study.

ACKNOWLEDGMENTS

I wish to thank all of the people who contributed in various ways to make this dissertation possible. First, the teachers in the Houston Independent School District who took the time to provide the foundational information for this study; I am grateful to you for sharing your words and experiences and I strive to present them in an impactful way.

I'd like to thank my faculty committee: Dr. David Berliner, Dr. Gustavo Fischman, and Dr. Audrey Amrein-Beardsley, for your wisdom and time. Audrey, you have served as an invaluable mentor and role model, and I am grateful for the opportunities you have provided me.

Without the support of my colleagues and friends, I would not have made it through our graduate program. You provided stability, and most importantly, you were my sounding boards and motivators. Dr. Lenay Dunn, thank you for blazing the path and providing most valuable advice. Rebecca Lish, you always know how to ask the questions I am trying to formulate and your wit is much appreciated. Jennifer Shea, you have more drive than anyone I have known, and an endless supply of inspirational quotes that mean more than you know. Erin Nolan, you are a dear friend, always there to discuss and empathize with life and school situations.

I am indebted to my family, for their love and support. Thank you to my sisters, for caring to ask about school and for always to picking up the restaurant tab for your college kid sister – even if I did drag it out for a decade. Thank you to my parents; you served as my first teachers and exemplified hard work from early

on. My Dad told me that no one can take degrees away from you, and earning them will only open more doors to opportunity. My mom always encouraged me to work hard and to do my homework, even when I wondered when I'd ever need to recite the amendments, memorize the periodic table of elements, or compute long division by hand. I am thankful for the hundreds of notes of encouragement she would stick in my school lunches, then mail throughout college and graduate school, usually accompanied by baked goods.

Finally, I thank my husband Rich, who has provided endless support, both financial and emotional, throughout my graduate school journey. Thank you for helping me to accomplish this goal and for the countless home improvement projects you have taken on to occupy your time. I look forward to the next chapter of our lives and uninterrupted football seasons.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x
CHAPTER	
1 Introduction.....	1
Value-Added Models	1
Purpose	4
Research Questions	5
Significance of the Study	6
Limitations.....	6
Overview of the Dissertation	8
2 History, Conceptual Framework, and Literature Review.....	10
The Foundation of Value-Added in Education	10
The Civil Rights Act.....	11
A Nation at Risk	15
EVAAS Beginnings.....	16
EVAAS Expanded.....	20
National Overview of Growth and Value-Added Model Use	22
Conceptual Framework.....	30
Empirical Studies on Value-Added	33
Reliability.....	33
EVAAS and Reliability	35

CHAPTER	Page
Validity.....	37
EVAAS and Validity.....	41
Formative Use.....	42
EVAAS and Formative Use.....	43
Intended Consequences and Claimed Benefits of EVAAS.....	46
Unintended Consequences.....	47
3 The Situation in Houston.....	50
Focus Group Discussions.....	51
Reliability.....	52
Validity.....	54
Formative Use.....	57
Unintended Consequences.....	58
High Stakes.....	58
Study of Terminated Teachers.....	58
Teacher A.....	62
Reliability.....	62
Validity.....	63
Teacher B.....	64
Reliability.....	64
Validity.....	65
Teacher C.....	66
Reliability.....	66

CHAPTER	Page
Validity	67
Teacher D	68
Reliability	68
Validity	69
Overall Findings	70
Reliability	70
Validity	70
Formative Use	71
Unintended Consequences	71
Final Verdicts	72
Merit for Dissertation Study	73
4 Methods	74
Survey Research Study	74
Mixed Methods Approach	76
Quantitative Data	77
Qualitative Data	78
Survey Instrument	79
Participants	81
Data Cleaning	83
Response Rate	84
Generalizability	85
Data Analyses	86

CHAPTER	Page
Validity	88
Role of the Researcher	90
5 Results	93
Demographics and Description of Sample	94
Reliability	97
Reliability across Grade Levels.....	99
Reliability across Subject Areas.....	100
Reliability across Student Characteristics.....	102
Validity	103
Content Validity.....	104
Criterion-related Evidence of Validity.....	105
Formative Use	109
Formative Use Support	115
Intended Consequences and Claimed Benefits of EVAAS	118
Unintended Consequences	123
Disincentives for Teaching Certain Groups of Students	124
Teacher Mobility Issues.....	125
Cheating and Teaching to the Test.....	126
Distrust, Competition, and Low Morale	129
Summary of Results	132
6 Findings and Conclusions	133
Summary of the Study.....	133

CHAPTER	Page
Overall Findings and Implications.....	135
Reliability.....	135
Reliability Implications	137
Validity.....	138
Validity Implications	140
Formative Use.....	141
Formative Use Implications	143
Intended Consequences	145
Intended Consequences Implications.....	145
Unintended Consequences	146
Unintended Consequences Implications	147
Cultural Consensus	147
Conclusions	149
Recommendations for Further Study.....	151
REFERENCES	155
APPENDIX	
A ARIZONA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPROVAL	167
B HOUSTON INDEPENDENT SCHOOL DISTRICT RESEARCH APPROVAL	169
C SURVEY PROTOCOL.....	172

APPENDIX	Page
D INTRODUCTION AND REMINDER EMAILS TO HOUSTON INDEPENDENT SCHOOL DISTRICT TEACHERS	178
E LIKERT-SCALE TABLE WITH PARTICIPANT RESPONSE PER ITEM.....	181
F CHI-SQUARE ANALYSES RESULTS FOR LIKERT-SCALE ITEMS.....	183
G CHI-SQUARE ANALYSES RESULTS FOR ALL OTHER CATEGORICAL ITEMS.....	186

LIST OF TABLES

Table	Page
1. Teacher A’s EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010).....	63
2. Teacher B’s EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010).....	65
3. Teacher C’s EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010).....	67
4. Teacher D’s EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010).....	69
5. EVAAS Data Usage by Teachers	113
6. EVAAS Training Session Attendance	116
7. Items Capturing Respondents’ Opinions about EVAAS Statements	119
8. Chi-square Analysis for Statement 1	121

LIST OF FIGURES

Figure	Page
1. Growth and Value-Added Model National Overview	24
2. Legislation Requiring Teacher Evaluation to be Linked to Student Growth or Achievement Data	25
3. A Teacher’s EVAAS Teacher Value-Added Report for 2010 and EVAAS Report for Teacher Reflection	45
4. Teacher C’s EVAAS Teacher Value-Added Report for 2010 and EVAAS Report for Teacher Reflection	61
5. Number of Years for Which Individual EVAAS Scores Were Received.....	94
6. Proportion of Grade Levels Ever Taught in the Houston Independent School District.....	95
7. Proportion of Subject Areas Ever Taught in the Houston Independent School District.....	96
8. When Teachers Typically Receive EVAAS Reports for Their Students.....	110

CHAPTER 1

Introduction

In recent years, the topic of teacher effectiveness and the need for strong evaluation systems has become a growing trend of national focus. Many states have or are in the process of reforming teacher evaluation procedures to account for teacher contributions to student achievement, specifically by using student test scores. This emphasis is a result of educational policies and incentives such as No Child Left Behind (NCLB), Race to the Top, and the Teacher Incentive Fund that highlight the importance of high quality teachers, and now require states to measurably demonstrate teachers' impact on student learning and achievement (U.S. Department of Education [USDOE], 2012). Impact is to be measured by student test score gains from year-to-year, and the growth or progress made is to represent the value-added by individual teachers (Newton, Darling-Hammond, Haertel & Thomas, 2010). The Value-Added Models (VAMs) being promoted and used are the statistical systems used to track the academic growth of students for teacher accountability.

Value-Added Models

VAMs measure student growth in achievement from one point in time to the next, using large-scaled standardized testing data. Scores are generated for each student, to compare their current year test scores to the past year, and also to compare student progress to that of their peers. These student growth scores are considered teachers' value-added, and value-added scores can be compared to purportedly distinguish effective from ineffective teachers. VAMs are growth

models, but the distinguishing difference between the two is that VAMs contain covariates or “blocking” variables that can control for the effects of external factors, so as to “better” determine the individual impact that a teacher, school, or district has on student learning.

The SAS Education Value-Added Assessment System (EVAAS) was one of the first VAMs created and is the most recognized and most widely-used model in the country today. The EVAAS technically consists of multiple models that are used to calculate value-added depending on the type and availability of test data. The most common EVAAS model is the multivariate response model (MRM) which is a multivariate model that contains multiple years of testing data for each individual student and vectors of random and fixed effects (Wright, White, Sanders & Rivers, 2010). The MRM is the most sophisticated and preferred model, given its ability to account for more than the less sophisticated models (e.g., the univariate response model [URM]). The URM model uses student scores in one grade/subject as the dependent variable, prior scores as independent variables, and either the teacher, school or district as the categorical variable (Wright et al., 2010). If a school or district does not have sufficient data for the MRM, the URM is used.

The EVAAS model is self-proclaimed to be “the most comprehensive and reliable” system available, better than the “other simplistic models found in the market today” (SAS, 2012a). As advertised, the system “provides valuable diagnostic information about [instructional] practices,” helps educators become more proactive and make more “sound instructional choices,” and helps teachers

use “resources more strategically to ensure that every student has the chance to succeed” (SAS, 2012a). The EVAAS, like most other VAMs, has been shown to be more accurate at analyzing student academic progress than traditional end-of-year “snapshot” or Adequate Yearly Progress (AYP) reports, and the EVAAS is probably the best or “least bad” VAM in existence (Amrein-Beardsley, 2008; Economic Policy Institute [EPI], 2010); however, critics raise concerns with the model’s validity, consistency, reliability, and lack of transparency (Amrein-Beardsley, 2008; Eckert & Dabrowski, 2010; Newton et al., 2010).

As well, the growing national focus on measuring teacher quality, most recently heightened by NCLB waiver application requirements (USDOE, 2012), has led to an increasing number of states and districts turning to for-profit models, such as the EVAAS, without fully understanding the intended and unintended consequences of model implementation. Despite widespread popularity of the EVAAS, however, no research has been done to examine how teachers and their practices are impacted by this methodology that professedly identifies effective and ineffective teachers. Even more disconcerting is that districts and states are tying consequences to the data generated from the EVAAS, entrusting the sophisticated methodologies to produce accurate, consistent, and reliable data.

Existing research on value-added, and the EVAAS model in particular, tends to be largely quantitative, focusing only on the data generated from the model. Lacking from the research base, are (mainly qualitative) studies about the relationship between value-added scores and the teaching qualities they are assumed to measure (Hill, Kapitula, & Umland, 2011), as well as analyses of how

these models actually impact the teaching profession. As a result, whether teachers use the data to reflect and improve their instruction and teaching strategies remains essentially unknown. To truly realize if the EVAAS works in practice as intended and provides the benefits advertised by SAS, it is necessary to bring the invaluable perspective and experiences of the teachers into the national conversation.

Since 2007, the Houston Independent School District (HISD) has contracted with the SAS Institute to measure student growth and progress using the EVAAS model, and specifically the MRM model just mentioned. HISD uses EVAAS as basis for a teacher merit pay program, and as part of their teacher evaluation system where EVAAS scores are used and can ultimately impact termination decisions. HISD was selected as the district for this study because they are using value-added data more than any other district or state in the country for high-stakes purposes (Corcoran, 2010; Harris, 2011; Mellon, 2010; Otterman, 2010; Papay, 2010). Greater details about HISD and their teacher evaluations are provided in Chapter 3.

Purpose

The purpose of this study was to gain an understanding of how teachers and their teaching practices have been impacted by the implementation and use of the EVAAS model to hold them accountable within HISD. The study was designed to investigate the reliability, validity and formative use of EVAAS data as experienced by the teachers. Additionally, the study investigated the intended consequences of the EVAAS, the benefits and outcomes as marketed and

promoted by SAS; as well as the unintended consequences occurring as a result of the EVAAS model implementation, many of which emerged as a result of analyzing teachers' reported experiences with the EVAAS in HISD. The intent of this study was also to bring the invaluable perspective of the teachers into the research and policymaking conversations, by examining the words and experiences of teachers.

Research Questions

By conducting this study, I sought to understand teacher experiences with the EVAAS as well as how the model impacted their teaching practices. The overarching research question was: What are the intended and unintended consequences, as experienced by HISD teachers, through the implementation and use of the EVAAS model? To research this question, I designed a large-scale electronic survey questionnaire that contained four different constructs with sub-questions regarding: (a) Reliability – Are EVAAS scores consistent over time? (2) Validity – Do EVAAS scores match other indicators of teacher quality? (c) Formative use and consequences – Do teachers use EVAAS data to inform their instruction? (d) Intended consequences and claimed benefits of EVAAS – Do teachers agree with EVAAS marketing claims and statements? I also included demographic questions pertaining to participants' years of teaching experience, subject area(s) taught, grade level(s) taught, and where they received teaching certification.

Significance of the Study

This study is one of the first to investigate how the EVAAS model works in practice. It exposes never heard before experiences, particularly in terms of the intended and unintended consequences of using the EVAAS model for high-stakes consequences in HISD. It also examines teachers' perceived realities of the model's advertised utility.

Beyond the scope of the EVAAS, this study also provides a glimpse of whether VAMs might produce desired outcomes and encourage best teacher practices. Though particular teachers and the district may represent an isolated case of using value-added for high-stakes consequences, this study nonetheless contributes valuable information for policymakers, researchers, and districts. This is information of which others should be aware and others should consider when implementing the EVAAS, or any value-added or growth model for teacher evaluation for that matter, as many of the issues that emerged through this study are not solely attributed to the EVAAS.

Limitations

This survey research study, like all studies, has its limitations. Only the data of EVAAS eligible teachers (social studies, science, math and English language arts/reading in grades 3-8) were included in the final analyses. But the respondents were not randomly selected. Rather, voluntary participants completed this electronic survey as distributed to all K-8 teachers within HISD. It could be argued that only those who had strong opinions on the matter of EVAAS were the teachers who responded to the survey, and this should be kept in mind when

reading the results. Nevertheless, there is still something to be learned from the experiences shared herein. Issues with generalizability will be discussed in greater detail in Chapter 4.

A second limitation is the strong union presence among the teachers who participated in this survey research study; however, I conducted chi-square analyses to examine this disproportionate representation, and results indicated that the responses of union teachers did not statistically differ from the responses of non-union teachers whatsoever. These analyses notably diminish anticipated concerns about whether the results of this study were biased by the strong participation of union teachers.

A third limitation pertains to my own bias given my role as the researcher. Through the explanation of the research methods, my intent is to convey credibility by describing the framework I used to design the survey constructs, as well as through the data themselves (Glaser & Strauss, 1967). Additionally, the rich description and detailed information provided by the voices of the teachers serve as evidence for my findings (Denzin, 1989; Gay & Airasian, 2003). Their data should speak for themselves, and increase my credibility as a researcher given they are left intact and only used to substantiate the major and minor findings throughout.

Due to these limitations, I employed particular validation and triangulation techniques, mainly to remain true to the data and keep these limitations in check. Specifically, I triangulated across the quantitative and qualitative responses from this study to support the study findings with solid evidence gathered from

multiple perspectives. I shared the findings with HISD teachers for member checking purposes, and they ultimately verified and helped to refine my final findings. And again, I used the detailed descriptions resident within the qualitative data to convey and evidence findings in the teacher respondents' words, not just my own. These strategies help to enhance the validity of the study (Creswell, 2003). These efforts along with additional information about my role as a researcher will be discussed more thoroughly in Chapter 4.

Overview of the Dissertation

In Chapter 2, I provide a historical analysis of value-added, from its inception in economics to today's use in teacher evaluations, and with a specific focus on the creation and evolution of the EVAAS model. I provide an overview of the value-added platform nationwide, as states have adapted to the changing federal teacher evaluation policies and mandates. A version of the national overview information on VAM usage is currently under review for publication (Collins & Amrein-Beardsley, under review). Additionally in this chapter, I explain the conceptual and analytical frameworks that informed and facilitated the design of this study. Finally, I review the empirical research and studies on key value-added issues.

In Chapter 3, I introduce the EVAAS situation in HISD and detail the preliminary research conducted in the district to merit this study. This includes information gathered from two focus group discussions and also the EVAAS data from four terminated HISD teachers. A similar but different version of Chapter 3 was published with my dissertation chair and co-author Dr. Audrey Amrein-

Beardsley (hereafter, *Dr. Amrein-Beardsley*) in *Educational Policy Analysis Archives* (Amrein-Beardsley & Collins, 2012).

In Chapter 4, I explain the mixed methods research approach that I used to conduct this study, as well as the data collection and cleaning processes. After discussing the analyses and generalizability of the results, I discuss the measures I took to ensure validity. I conclude the chapter with an explanation of my role as a researcher and the limitations of the study.

In Chapter 5, I discuss the main results for each survey construct and the overall results. Using the words and experiences of the teachers, I allow their voices to be heard through mine. I use charts and figures to present both quantitative and qualitative results, and I summarize the overall themes from this study.

Finally, in Chapter 6, I conclude the dissertation by summarizing the study, and then discussing the findings inferred from the results and additionally supported by literature. Then I discuss the implications from the study as it relates to both local and national educational policy and provide recommendations for further research on this topic.

CHAPTER 2

History, Conceptual Framework, and Literature Review

In this chapter, I explain the evolution of value-added in education, which includes the beginning of econometric models used in education, the history of education accountability that led to the demand for value-added measures, the history of the EVAAS model, and information about the current national use of growth and VAMs. Then I describe the conceptual framework that guided this study, as well as the analytical framework I used to design and execute the survey research study. Finally, I highlight the key issues pertaining to value-added and value-added implementation as discussed in the research.

The Foundation of Value-Added in Education

The VAMs used in education stem from the field of economics. In a very basic economic form, value-added is the calculated difference between a set of inputs and outputs predetermined in an econometric model. In manufacturing firms, for example, value-added is the difference between the price of a finished product and the cost of production, including the cost of resources and parts used in the production process (Fincher, 1985). Value-added represents the quality of goods with the value increased by advanced levels of technology and skill in production, assuming the production process requires some type of “human agency” to add value to the end-product (Saunders, 1999, p. 236).

Applying this to the education setting, value-added represents the value that a teacher, as the input, adds to or detracts from student learning, as the output. In other words, the value-added (or detracted) is synonymous with the student

achievement growth (or decline) year-to-year. Though variations of VAMs exist with different inputs or variables and controls included in the models, the output is always measured by student growth on some type of large-scaled standardized test. VAMs are now used as the key component for teacher accountability across the United States today, particularly given the latest iteration of the nation's educational accountability movement (Ravitch, 2012a).

In the next section I describe the foundation of the educational accountability movement in the 1960s as well as introduce the work of Eric Hanushek, who used some of the earliest econometrics VAMs to look at educational inputs and outputs for school and teacher accountability starting in the 1970s.

The Civil Rights Act. The Civil Rights Act of 1964 initiated education accountability in the United States. Section 402 of the Civil Rights Act required a national report of the equal educational opportunities available for all individuals after federal courts had discovered inequitable distribution of educational funding and resources to schools of primarily non-white minority students (Coleman, 1966). The inequities were discovered by sociologist James Coleman, author of the Equality of Educational Opportunity report, otherwise known as the Coleman Report, in 1966. Coleman's cross-sectional study relied on data derived via the National Center for Educational Statistics (NCES), which included data representing over one half of a million students from more than 3,000 schools over the course of one school year. The study found many inequities across schools including class sizes, student achievement levels, school quality,

availability of school resources, and teacher quality as measured by the education levels and training of teachers (Coleman, 1966). Teacher quality was found to have the greatest impact on student achievement compared to all other school-related factors, however.

According to Hanushek (1979), the Coleman Report first introduced the critical notion of how school inputs impact student achievement and evidenced that variation in teacher quality had a cumulative effect on students as they progressed through school. Noting the inequalities highlighted by the Coleman Report, Hanushek (1971) explained the difficulty in improving the equitable distribution of resources because there was so much unknown about the relationship between educational inputs (e.g., teachers, curricula, peer students, facilities) and outputs (e.g., multidimensional factors composed of students' achievement and attitudinal changes). Up to the 1970s, there had been little to no historical data available at the individual student-level. Instead there was a societal emphasis on educational inputs instead of outputs, meaning relatively little was known about how schools and teachers actually affected the education process. For instance, it was assumed that tenure and advanced college education resulted in more effective teachers and increased student learning; however, no studies had then been done to evaluate these hypotheses (Hanushek, 1971).

To further investigate the relationship between inputs and outputs, Hanushek (1970) conducted a study to look at three fundamental questions surrounding the educational process: whether teachers mattered in the learning process, how efficiently schools were operating, and what characteristics of

teachers and classrooms were important. In a school district in southern California, he tracked students from first through third grade to examine the relationship between school system inputs and outputs “as measured by achievement scores and attitudinal change” (Hanushek, 1970, p. IV). His model used data from each student’s education level (via first grade Stanford Achievement Test scores) so that it was possible to determine the value-added by measuring gains in achievement during the second and third grades. Other inputs in Hanushek’s model included socioeconomic status (it was generally accepted to highly correlate with family educational inputs), peer classmates’ influence, innate abilities (e.g., IQ scores), and school influences, which were based on his hypothesis that tenure and further schooling equated to higher quality teaching and that class assignments had a beneficial effect on education.

Hanushek (1970) concluded that teachers do not appear to impact the learning of Mexican American students, but found that significant differences in the performance of white children were dependent on the teacher, regardless of the student’s socioeconomic status. His findings essentially contradicted the Coleman Report; however, Hanushek was unable to identify characteristics of effective teachers so the information was declared unhelpful to administrators.

Several years later, Hanushek (1979) revisited the economic notion of inputs and outputs in education, this time looking at production function models. In a traditional economic or manufacturing setting, two production processes applying the same inputs should result in the same outputs, and any differences would indicate inefficiencies. In education however, students having the same

inputs (e.g., classroom, teacher, school) can most certainly yield different achievement outputs which are not necessarily issues of inefficiency. Despite the inability of production function to measure inefficiencies in education, however, Hanushek (1979) found the model to be useful in providing information on characteristics of teaching that could be replicated in hopes of reaching desirable outcomes in student achievement.

Although other 1970s studies addressed the value of education or the measure of education quality (Boudon, 1974; Duncan, Featherman & Duncan, 1972; Jenks et al., 1972; Sewell & Hauser, 1975), they failed to adequately capture the value that education added. Hanushek (1979) attributed such research inadequacies to model inputs that were limited by data availability, and the models themselves relying on available data instead of variables that might be more telling or desirable. Additionally, findings by earlier researchers (see above) provided inconsistent and unverifiable results due to differing samples, the varied levels of data collected (school-level versus individual student- and teacher-levels), and the various types of analytical models used (Brophy, 1973; Hanushek, 1979).

Hanushek's econometric model that he first used in the 1970 study in southern California was one of the first "value-added" models to be derived from conceptual needs and not based on data availability. Hanushek's model was also one of the first to include inputs with cumulative influence (e.g., family background influences, classroom or peer influence, and school influence) on student achievement, which he believed had lasting impacts on student

achievement year to year (Hanushek, 1979). His foundational studies of value-added measures, particularly to measure teacher inputs, were timely as education reform at the national level was about to begin focusing more on teacher quality.

A nation at risk. The 1980s represented the pioneer days of educational reform efforts specifically focused on test-based accountability and consequences for teachers and school systems (Koretz, 1996). *A Nation at Risk* was released in 1983 by the National Commission on Excellence in Education (NCEE) under the Reagan administration. This report spearheaded the emphasis to restructure the American educational system to produce an educated workforce to maintain the country's economic integrity and competitiveness, which were both reportedly in jeopardy as a result of our failing education system (NCEE, 1983). As a result of this report, the national focus turned toward teachers, and their potential to influence and educate the future workforce.

Ambach, for example, referenced the 1980s as a period when the nation began both measurement-driven instruction and measurement-driven educational policy (as cited in Koretz, 1996), with nearly every single state developing their own testing policies during this time. Reform efforts focused primarily on increasing student learning as well as the kind of teaching that was necessary to facilitate high quality learning (Center on Organization and Restructuring of Schools, 1995). Teacher qualifications and credentials were the criteria on which school officials concentrated as they began to look at teacher accountability (Meyer, 1997). While the rest of the country was grappling with *A Nation at Risk*

and how to increase both teacher quality and the educational system as a whole, the makings of EVAAS were starting to take form in the state of Tennessee.

EVAAS beginnings. Dr. William Sanders (hereafter, *Dr. Sanders*) spent the early parts of his career as an adjunct statistics professor, working at the University of Tennessee's school of agriculture (Hill, 2000). Dr. Sanders' crossover into educational statistics derived from agricultural statistician Charles Henderson, who had applied statistics to genetic trends and breeding methods for livestock (Kennedy, 1991). Henderson (1973) created mixed model selection, which means that the subjects are treated as a random sample of variables with unknown means. His mixed model selection allowed for certain variables of choice to be fixed and others to be random. Most significantly, the model provided a technique to determine if a selection of variables would produce bias estimates to help identify best linear unbiased predictions (BLUP) (Henderson, 1973). BLUP is used in mixed models to estimate random effects.

During Tennessee's educational reform efforts in the 1980s, Dr. Sanders was teaching an advanced-level statistics course at the University of Tennessee, and used the example of linking student test scores back to their teachers (Gabriel & Lester, 2012). As this was a timely educational issue, Dr. Sanders and his colleagues continued working on exploratory statistical mixed-model methodologies to try to avoid previous issues with student achievement data such as missing data, different teachers and teaching assignments year-to-year, regression to the mean, and student mobility (Sanders & Horn, 1994).

Sanders and McLean developed a system of analyses based on Henderson's mixed-model methodology that was first used with three years of longitudinal student data from Knox County school district in Tennessee. They found strong correlations between teacher effects as determined by student data and supervisor evaluations, and also differences among schools and teachers and their unique impacts on indicators of student learning (Sanders & Horn, 1994). To verify their results, Sanders and McLean applied the same model to student data from two other districts in Tennessee. Though the findings validated their original results, it would be several years before the model gained state-wide attention (Sanders & Horn, 1994).

In 1989, a lawsuit filed by a group of small Tennessee school districts claimed that the state was violating their constitutional rights by not providing equitable funding across all districts to provide equal educational opportunities for all students (Sanders & Horn, 1998). Policymakers wanted to increase taxes to generate extra funding for the poorer districts, but this idea was not well accepted by the public (Ceperley & Reel, 1997). To generate the necessary tax money, policymakers turned to the financial support of businesses that required several forms of accountability before they were willing to provide money for the schools (Ceperley & Reel, 1997). Businesses wanted principals to be accountable for executing performance contracts with consequences for those who failed to fulfill contract obligations. They also wanted classroom accountability with sound "evidence that dropout rates, promotion rates, proficiency test passage rates, and student achievement were improving from year to year" (Ceperley & Reel, 1997,

p. 134). The business requests were met in the creation of Tennessee's Education Improvement Act (EIA) of 1992, which simultaneously called for an increase in state education funding and demanded a stronger accountability condition to ensure money spent was actually improving student academic achievement (Sanders & Horn, 1998).

By this time, Dr. Sanders had expanded on the original model and was able to provide school system effects on the academic progress of students in grade levels 3-8 using scores from the norm-referenced Tennessee Comprehensive Assessment Program (TCAP) (Sanders & Horn, 1994). The focus of the accountability movement in Tennessee was centered on the "*product* of educational experience rather than the *process* by which it was to be achieved" (Sanders & Horn, 1994, p.300). Therefore, Dr. Sanders' outcomes-based assessment system emerged as the perfect tool to measure student achievement and provide the precise accountability system demanded by the EIA legislation (Ceperly & Reel, 1997; Sanders & Horn, 1998).

The Tennessee Value-Added Assessment System (TVAAS), also called "Sanders's Model," became officially recognized by state legislation and was included in the 1992 EIA plan for assessing progress and providing information about the contribution teachers, schools, and school systems made to student learning gains (Sanders & Horn, 1994; Tucker & Stronge, 2005). The TVAAS focused on achievement gains from all students year-to-year, which meant (at that time, although this continues today) the most successful schools were those that increased learning opportunities for all types of students (Sanders & Horn, 1994).

This included advanced students as well as those who entered classrooms with achievement levels below grade level.

In 1993, Tennessee began using reports generated from TVAAS to provide “teachers and administrators with estimates of teacher effectiveness” (Sanders & Horn, 1998, p. 248). Although Tennessee legislation required the TVAAS reports to be used as a component of the teacher evaluation system, the proportion of evaluation to be based on the data was left to individual choice by school districts. Sanders & Horn (1998) cautioned that student achievement data from the reports were not to be the only source of data used in a teacher’s evaluation.

Nonetheless, the Tennessee State Board of Education contracted with Dr. Sanders through the University of Tennessee for services related to TVAAS from 1992 until 1999 (Morgan, 2004). In 2000, Dr. Sanders retired from the University of Tennessee and took the TVAAS to SAS Institute, Inc. There, he changed the name of the TVAAS model to Education Value Added Assessment System (EVAAS) (SAS, 2011). At the same time Dr. Sanders left the state, Tennessee introduced the Framework for Evaluation and Professional Growth. After several years of re-evaluating the teacher evaluation process, improving student performance became the centerfold focus within teacher evaluation in Tennessee (Tucker & Stronge, 2005). The State Board of Education continued contracting with Dr. Sanders through SAS to provide TVAAS evaluations for the teachers in Tennessee.

EVAAS expanded. Meanwhile at the national level, the NCLB Act of 2001 was released by congress. NCLB emphasized measurable student achievement goals, which resulted in states reporting AYP on student assessment scores. In 2005, the U.S. Secretary of Education announced a set of new guidelines, however, which mandated states to assess all students annually in grade levels 3-8 and in high school with emphases on the core subject areas of reading and mathematics (Ed.gov, 2008). States were also required to provide results by student subgroups and improve teacher quality (Ed.gov, 2008).

Later that year, the federal government started a growth model pilot program for qualified states to use growth-based accountability models as an alternative option to see if they would be more effective in measuring and increasing student achievement than traditional AYP analyses and reports. Not surprisingly, Tennessee and North Carolina were the first two states approved to participate in the pilot (USDOE, 2008), most likely because of pre-existing contracts with SAS and their state-wide use of the EVAAS. Ohio and Pennsylvania also contracted with SAS (SAS, 2011) and eventually participated in the pilot program as well (USDOE, 2008). In total, 15 states were approved and funded to participate in the pilot program between 2005 and 2010 (Ed.gov, 2008).

After the federal growth and value-added model pilot ended, the USDOE released a report which indicated that the impact of such models on student achievement was minimal (Carey & Manwaring, 2011). However, the minimal impact was attributed to different models that relied on different years of data, varied degrees of difficulty among state standards and tests, and most importantly,

varied interpretations of growth or value-added (Carey & Manwaring, 2011). Even though the pilot ended, and despite meek pilot findings, federal incentives (e.g., NCLB waivers and Race to the Top) continue to incite states to pursue reform efforts to measure student growth and teachers' contributions toward that growth (Ravitch, 2012a).

Around the same time the pilot ended, and also bolstering this movement, the New Teacher Project produced "The Widget Effect," a report explaining that once again, as a nation, we have failed in our education system (Weisberg, Sexton, Mulhern & Keeling, 2009). Much like with *A Nation at Risk*, it seemed our country faced yet another "manufactured crisis" (Berliner & Biddle, 1995) with faulty schools full of mediocre teachers. This time the tactics focused entirely on teachers, those inherently responsible and proven to have the most significant in-school influence on student learning (Sanders & Rivers, 1996).

The Widget report also highlighted the inability to distinguish good teachers from bad, calling teachers "widgets" as a result of our nation's faulty teacher evaluation systems which currently rate, on average, 99% of all teachers as effective and 1% the inverse (Weisberg et al., 2009). In response to this report and others with similar accord (Corcoran, 2010; Goldhaber & Hansen, 2010; Hanushek, 2011), the race was on so-to-speak, for a new, more objective, discerning teacher evaluation system that could properly identify effective, average, and ineffective teachers.

National Overview of Growth and Value-Added Model Use

The federal accountability efforts further pushed states and districts away from focusing on student-level to teacher-level accountability and toward rewarding and punishing measurably effective and ineffective teachers respectively. Likewise, econometricians and statisticians (e.g., SAS) increasingly claimed (and continue to claim) that they could help states and districts reliably and precisely identify good and bad teachers using student test scores, and help them do this well enough to support highly consequential decisions about the teachers identified (e.g., merit pay, denial or removal of tenure, teacher termination). As a result, the majority of states are now planning if not already using growth or VAMs to track the academic growth of students, and to attribute such changes to the students' teachers.

Given the recent and growing focus on growth models, however, it was difficult to find a resource that provided an overview of what each state is currently doing to measure teacher effectiveness. To gather this information as part of my doctoral work (i.e., via my research internship) and to inform this dissertation project, I (along with Dr. Amrein-Beardsley) examined what all 50 states and Washington D. C. (hereafter, *D.C.*) are currently doing to measure teacher effectiveness via growth and VAMs (Collins & Amrein-Beardsley, under review). I collected information from state department of education personnel in charge of each state's initiatives in this area via phone interviews, electronic surveys, or website research. Although this is a rapidly changing arena, especially

given recent requirements for NCLB waivers (USDOE, 2012), this study provides the most inclusive report on national growth and VAM use available to date¹.

Currently, 40 states and D.C. are using, piloting or developing, some type of growth or VAM. The Student Growth Percentiles (SGP) model (also commonly recognized as the Colorado Growth Model) is used or piloted by 12 states (24%);² eight states and D.C. (18%) are using or piloting a VAM (including the EVAAS);³ Missouri is piloting both a growth and VAM; and Delaware is using a value table model. Additionally, 18 states (35%)⁴ indicated they are currently developing a model to be used statewide but did not specify a particular model. In three states (6%),⁵ growth or value-added use is locally controlled at the

¹ State data were collected and verified between July – December, 2011. Given the volatility of the growth and value-added model climate, it is possible this information may have changed, however.

² Arizona, Colorado, Hawaii, Indiana, Massachusetts, Mississippi, Nevada, New Jersey, New York, Rhode Island, Virginia, West Virginia

³ Florida, Louisiana, North Carolina, Ohio, Oklahoma, Pennsylvania, Tennessee, Wisconsin

⁴ Arkansas, Connecticut, Georgia, Iowa, Idaho, Illinois, Kansas, Kentucky, Maine, Maryland, Michigan, New Mexico, Oregon, South Carolina, Texas, Utah, Washington, Wyoming

⁵ California, Minnesota, Nebraska

district-level, and seven states (14%)⁶ indicated they do not have plans to develop a statewide growth or VAM for evaluating teacher effectiveness, although some are using such measures to evaluate school effectiveness (see Figure 1).



Figure 1. Growth and value-added model national overview.

In addition, 30 states and D.C. (61%) now have legislation or regulations that require student achievement, growth, or value-added data to be used in the evaluation of teacher effectiveness (see Figure 2).

⁶ Alabama, Alaska, Montana, New Hampshire, North Dakota, South Dakota, Vermont



Figure 2. Legislation requiring teacher evaluation to be linked to student growth or achievement data.

In terms of high-stakes consequences, nine states and D.C. (20%)⁷ use (or plan to use) growth or value-added output to differentiate levels of teacher compensation, award merit pay, or make pay-for-performance decisions. Ten states and D.C. (22%)⁸ tie (or plan to tie) teacher tenure decisions to such output, and nine states and D.C. (20%)⁹ use (or are planning to use) these data to make teacher termination decisions.

⁷ Florida, Indiana, Maine, Maryland, New York, North Carolina, South Carolina, Tennessee, Virginia

⁸ Florida, Hawaii, Indiana, Kentucky, Louisiana, Michigan, Minnesota, New York, Rhode Island, Tennessee

⁹ Hawaii, Kentucky, Louisiana, Maryland, Michigan, Minnesota, New York, Rhode Island, Tennessee

Fourteen states (27%)¹⁰ indicated that their teacher evaluations are (or will be) based on multiple measures of student achievement data (e.g., supplemental testing, student work portfolios), which is in line with the field standards developed by the prominent national associations on educational measurement and testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). These standards note most importantly that high-stakes decisions “should not be made on the basis of test scores alone. Other relevant information should be taken into account to enhance the overall validity of such decisions” (AERA, APA & NCME, 1999). These 14 states seem to be constructing more holistic teacher evaluation systems that follow the professional field standards.

South Carolina, for example, uses student work samples collected from all teachers’ classrooms, and not just the teachers who teach the core curricular areas typically assessed using these models. Maryland is currently using local or school-level data to contribute to individual student growth calculations. Kentucky is supplementing the state’s annual test data with interim assessment and student portfolio data.

¹⁰ Delaware, Georgia, Hawaii, Kentucky, Maryland, Nevada, New Hampshire, New Jersey, North Carolina, South Carolina, Virginia, Washington, West Virginia, Wisconsin

With regard to the tests used for calculating growth or value-added, 100% of the 22 states and D.C. that are currently using or piloting the models use their state standardized tests in mathematics and English/language arts for grade levels 4-8. Nine states (18%)¹¹ indicated that they evaluate (or plan to evaluate) teacher effectiveness at the high school level, using end-of-course exams. South Carolina is the only state evaluating early childhood teachers, using the Northwest Evaluation Association Measures of Academic Progress (NWEA MAP) for grades K-3, although Wisconsin also has plans to evaluate K-3 teachers as well.

In addition to questions about the logistics of growth and VAMs, I asked state department of education personnel to share their perceptions regarding the strengths and weaknesses of the models used in their states. Regardless of the type of model used, almost all states expressed concerns about assessing student progress for teachers of non-tested grades and subject areas. The issue of fairness was most troublesome (see also Darling-Hammond, Amrein-Beardsley, Haertel & Rothstein, 2012; Glazerman et al., 2011). Recall that 100% of the states that calculate (or with plans to calculate) growth use (or plan to use) their large-scaled, standardized test score data, predominantly collected in grades 4-8 in the core subject areas of mathematics and English/language arts. This means that a large majority of teachers are ineligible for growth or value-added evaluations. As one state representative noted, entire buildings can lack these types of scores (e.g.,

¹¹ Florida, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Utah, Virginia, Wisconsin

early childhood/primary and high schools campuses), making evaluative comparisons nearly impossible, not to mention unfair. In fact, approximately 70% of all teachers nationwide cannot be evaluated, rewarded, or penalized using these models or the data to be derived from the models (Harris, 2011).

Several states also expressed concerns with validity as it relates to their state standardized assessment systems. Carey and Manwaring (2011) point out that “growth models, like all test-based measurement systems, are only as good as the test on which they rely. Many of the standardized tests used in K-12 education are inadequate, and their flaws can be magnified by growth calculations” (p. 10). Although some states boasted about the perceived strengths of their state standardized testing systems, others expressed concerns about whether the tests designed to measure growth or value-added accurately captured teacher effectiveness. Some state representatives mentioned that weak tests weakened validity, whereas states with self-reported “stronger tests” felt their tests strengthened validity. Other states mentioned validity-related challenges with ensuring proper linkages between students and their teachers of record for accurate data analyses, reporting, and recording. Several state representatives also pointed out that teaching consists of multifaceted, collective, and cumulative efforts that occur all-year long, whereas these models evaluate everything teachers do using only one test every year, or two tests when growth is measured from point x to y under the instruction of different teachers. To counter this, state respondents in particular who use (or plan to use) multiple measures of student

achievement, stressed the numerous ways teacher effectiveness would be captured and measured beyond state standardized tests alone.

When I asked state personnel about the use of growth and VAM data for formative purposes, not one representative from the 22 states and D.C. that are currently using (or piloting) models (45%) was aware of a state plan for using the data for formative purposes. According to respondents, the sophistication and complexity of the statistics used in states' growth and VAMs make it difficult to clearly explain the process and derived data to virtually anyone, including teachers and administrators, unless the creators of the models themselves conduct the explanations.

For those personnel representing states using the EVAAS model¹², they expressed the common concern that the model is proprietary, and although SAS has been willing to share many of the workings of the model with states, there are as in the words of one respondent, "lingering concerns about the transparency of the model." As another state coordinator expanded "the calculations of EVAAS results are not easily explained or visible" and this in itself prohibits the accessibility and usability of the EVAAS system. Respondents, however, noted as a strength that the model was able to show growth for all achievement levels of students, from the lowest to the highest performing students. Like with the SGP, this provides information on groups' of students or individuals' progress and their potential "to catch up." One state also saw predictive value in the EVAAS, which allowed the use of student-level data to predict future achievement, largely to

¹² New York, Ohio, Pennsylvania, South Carolina, Tennessee

inform the placement of students in courses with appropriate levels of rigor. It is important to note here, though, that whether predictions or the appropriateness of the placements made are actually verified is a concern (see also Amrein-Beardsley, 2008). Additionally, personnel representing states using the EVAAS appreciated that the model is able to account for students with missing test score data “in a sophisticated way.”

Regardless, growth and VAMs play a progressively more important role in teacher evaluations across the country, and yet no research has included teacher input regarding the models increasingly being used to evaluate them. To better understand this VAM trend and the consequences associated with using such models to evaluate teachers, I will describe the conceptual framework I used to guide this study and review the empirical research and studies that highlight the most significant issues with VAMs, again as framed using my analytical framework.

Conceptual Framework

My conceptual framework for this study borrows from cultural consensus theory (Romney, Weller & Batchelder, 1986) in that value-added accountability has become commonly and increasingly accepted by the general public as the means to improve teacher quality, albeit largely influenced by federal policy. Cultural consensus theory describes when trying to understand a phenomenon, we would expect those within the population of concern to have the answers that most closely represent the truth (Romney et al., 1986). Currently, value-added is seen and largely accepted by the general public as the logical tool for improving

teacher quality by identifying and eliminating ineffective teachers from the field. While some view value-added as “a measure of educational effectiveness that promises to revolutionize education” (Stone, 1999, p. 240), at minimum, the cultural consensus is that value-added is “good enough” to move forward, even if concerns remain (see Amrein-Beardsley, 2012).

There also appears to be cultural consensus that quality teachers are important and beneficial to students’ learning, yet how teacher effectiveness is defined or measured is less clear and lacks the same level of consensus, especially among teachers. This scenario presents an opportunity for further explanation and exploration, and teachers need to be a part of this learning opportunity. In describing cultural consensus theory, Romney et al. (1986) used a tennis example where a set of tennis questions were asked to two sets of informants: tennis players and non-tennis players. The authors explained:

We would expect that the tennis players would agree more among themselves as to the answers to the questions than would non-tennis players. Players with complete knowledge about the game would answer questions correctly with identical answers or maximal consensus, while players with little knowledge of the game would not. (p. 316)

Applying this to teacher effectiveness and value-added, we would expect more consensus on the topic of teacher quality among teachers – the individuals who have the “complete knowledge” and are actually evaluated by and purportedly using value-added – than among non-teachers. Romney et al. (1986) explained

that more importance should be placed on the responses from the knowledgeable informants than the less knowledgeable informants. In other words, we should be learning from and listening to teachers, instead of assuming that VAMs are “good enough” for high-stakes use.

Using this as my conceptual framework, I created an analytical framework to design my survey research study to explain the reality of this value-added social phenomenon from the perspective of teachers who have been evaluated by the EVAAS in HISD. The analytical framework helped me design and organize the key constructs within the survey, and was formulated using the “Standards for Educational and Psychological Testing” as designated by the leading national associations in educational measurement and testing (AERA, APA & NCME, 1999). The Standards include criteria to evaluate tests and the use of tests, and were designed for use by professional test developers and users (AERA, APA & NCME, 1999, p. 3). These criteria allowed me to investigate whether such recommendations were evident within the EVAAS model itself as well as the implementation and use of EVAAS within HISD.

As such, the constructs within my survey included questions pertaining to reliability, validity, formative use, and the intended consequences and claimed benefits of EVAAS. Unintended consequences also emerged via analyses of, in particular, respondents’ open-ended feedback. The issues I investigated are described in more detail throughout the next section, and the order in which I discuss the issues aligns with the analytical framework I used to frame my constructs and research questions of interest.

Empirical Studies on Value-Added

The debate in the research field over the merit and trustworthiness of the emerging growth and VAMs is noteworthy. In this section, I discuss many of the debated topics, including, as stated, issues with reliability, validity, formative use of growth and VAMs as well as the intended consequences and claimed benefits of EVAAS. Additionally I discuss whatever unintended consequences have begun to emerge via the literature related to the implementation of VAMs, and those specifically related to the implementation of EVAAS. These issues are organized and discussed in the same order in which they are presented in my survey protocol detailed in Chapter 4.

Reliability. The general meaning of reliability is to evidence trustworthiness or dependability. In measurement, reliability is “the degree to which a test consistently measures whatever it measures” (Gay, 1996, p. 145). Assuming that teaching practices remain relatively constant beyond the first few years of teachers’ careers (Baker et al., 2010; Harris, 2011), it would be expected that reliable growth and VAMs would consistently classify teachers as effective or ineffective from year-to-year, regardless of their teaching assignment and students in their classroom. Research has found, however, that not only are teachers classified differently year-to year using the same models (Baker et al., 2010; Corcoran, 2010; EPI, 2010), but using different VAMs can also yield strikingly different results with the same data (National Research Council [NRC], 2010; Newton et al., 2010; Rothstein, 2010). Switching the tests and measurements, or even shifting the timing between the pre and post-test

administrations also result in different or unreliable teacher classifications (Papay, 2010).

In attempt to combat some of the reliability issues, researchers recommend that two or more years of value-added data should be used to make decisions using value-added data (Assessment and Accountability Comprehensive Center [AACC], 2011; Brophy, 1973; Harris, 2011; Koedel & Betts, 2009), yet even that can result in a 25% chance of error, leading to the misclassification of teachers within constructed effectiveness levels (Otterman, 2010; Schochet & Chiang, 2010). For example, a Mathematica study found that even with 10 years of teacher data, the misclassification error was still 12% (Schochet & Chiang, 2010). Although error is inherent in statistical models, this is still a sizeable concern as states and districts are using limited (and faulty) data to make consequential decisions despite such statistical issues caused by error. As Harris (2011) explained, “When measures are used to make decisions, the statistical errors in the measures can result in decision errors” (p. 103) where certain teachers or schools can be systematically favored over others.

This seems to be the number one issue plaguing the practicality of VAM use to date. Many researchers argue that this occurs because student learning gains are impacted by demographic information (i.e., bias) and, hence, dependent on classroom effects beyond the teachers’ control. In other words, external factors such as socioeconomic status, home life, physical and mental health, motivation, behavior, etc. impact student test performance and cannot be controlled by teachers. Yet whether VAMs can control for demographics sufficiently is one of

the most hotly debated issues pertaining to value-added analysis and teacher classifications (Braun, 2005; Raudenbush & Bryk, 2002; Kupermintz, 2003; Newton et al., 2010). This issue is heightened by the nonrandom assignment of students to classrooms (Braun, 2004; Kupermintz, 2003).

EVAAS and reliability. To counter this, the EVAAS model is “continually refined to combine features of random and fixed effects models to increase precision while reducing measurement error and bias” (SAS, 2012b). Although some error is evident, Dr. Sanders claims that EVAAS reduces the chance of misclassifying teachers (as cited in Kupermintz, 2003), though he has not yet provided transparent evidence that the model reliably and accurately classifies teachers year-to-year (Amrein-Beardsley, 2008; Kupermintz, 2003). In fact, in a statement paper in response to EVAAS criticisms which included reliability concerns, Sanders and Wright (2008) specifically (and possibly deliberately) avoided addressing the reliability concerns others have raised about the EVAAS model and its (un)reliable teacher classifications.

Another contested EVAAS claim is the ability for the model to control for, or “block” external, student factors, from classroom composition to socioeconomic status, to isolate and examine only the effect that a teacher, school, or district has had on student learning (Sanders & Horn, 1998). According to Sanders and Horn (1998), students serve as their own control in EVAAS calculations, implying that demographic factors remain constant over time, which eliminates the need to adjust for such factors on an annual basis, and eliminates the need for the random assignment of students to classrooms. However, as

Kupermintz (2003) explained, “blocks” were created in controlled experiments that required random assignment of students for verification. Additionally, even though a student’s socioeconomic status may remain relatively consistent year-to-year, factors related to a student’s socioeconomic status (e.g., out-of-school learning opportunities as well as illnesses, familial or parental circumstances, troubles with law, sustained employment, safety and neighborhood considerations, etc.) will not be the same year-to-year and can cause fluctuations in year-to-year performance that are not constant.

Teacher mobility also compounds the issue of reliability, yet SAS states that it can address issues of teacher mobility as well (SAS, 2012c). Although EVAAS developers claim that teachers who move from one environment to another will continue to be as effective as in their previous teaching assignment, and they will be classified the same by their EVAAS scores (LeClaire, 2011), I was informed by a SAS employee that this statement would not apply to a scenario such as a fourth grade math teacher moving to teach fifth grade English/language arts classroom. Rather, this would make sense in a scenario where a math teacher moving grade levels or classrooms would receive consistent EVAAS scores regardless of the students in the classroom (J. White, personal communication, April 14, 2012). Evidence suggests that neither scenario is necessarily true (Amrein-Beardsley & Collins, 2012).

Despite the statements made by SAS, teacher mobility across grades and elementary subject levels is a very common reality and is also prone to have an impact on EVAAS scores. Such factors reduce the likelihood of teachers having

more than two consistent years of EVAAS data on which to base decisions.

Whether consistency across these varying teaching situations should be expected is debatable, but SAS has made such claims public, so this stands as a point for further research.

Validity. Validity is “the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment” (Messick, 1995, p. 741). Whether the meaning and resulting actions derived from the scores are understood across people, settings, or contexts is the test of validity (Cronbach, 1971; Messick, 1995). According to the leading national associations on measurement and testing, validity is “the most fundamental consideration in developing and evaluating tests” (AERA, APA & NCME, 1999, p. 9).

Though validity does not evaluate the test itself (Messick, 1980), there are numerous validity concerns with the proposed uses and application of VAMs and with the interpretation of value-added output. First, content-related evidence of validity describes the extent to which an assessment measures what it is intended to measure (Gay, 1996). In the case of value-added, student test scores derived from large-scaled standardized tests – which for myriads of reasons do not capture everything a student has learned in one year (Popham, 1999) – are used to evaluate teacher effectiveness. This issue is amplified when the assessments used to calculate value-added are not aligned with curriculum, although this is now less of an issue than it once was with the increase in criterion- versus norm-referenced tests post-NCLB (Darling-Hammond, 2007).

Nonetheless, concerns still exist, especially in districts and states that continue to use national norm-referenced tests (e.g., in the primary grades) along with criterion-referenced tests in determining value-added scores. Norm-referenced tests do not necessarily align with curriculum or depict understanding or mastery of a concept as well as criterion-referenced tests. In addition, tests that do not contain enough stretch to capture the growth of the very high-performing or gifted students is also of concern in terms of content-related evidence of validity (Amrein-Beardsley & Collins, 2012; Wright, Horn & Sanders, 1997).

Criterion-related evidence of validity, which includes both concurrent- and predictive-evidence of validity, is the degree to which criteria from at least two different tests or measurements correlate, and hence they are deemed to measure the same behavior(s) (Messick, 1980). Concurrent validity implies that the scores on a test are related to scores on another pre-established test, whereas predictive validity is the use of one test or measure to predict scores on another test or measure in a future situation (Gay, 1996). With regard to value-added, criterion-related evidence of validity is the correlation coefficient of concern among VAMs and other measures of teacher quality or effectiveness, namely in terms of concurrency.

Concurrent-evidence of validity has been a priority topic in some of the latest value-added research debates and discussions (H. Braun, P. Goldschmidt, D. McCaffrey, R. Lissitz, personal communication, April 16, 2012). These and other researchers (see, for example, Hill et al., 2011; Kane & Staiger, 2012; Sass & Harris, 2012) are actively discussing the relationship between value-added

measures and other measures assumed to evaluate the teacher effectiveness or quality construct as well, such as principal and peer evaluations.

For example, there has been much discussion about whether there is a correlation among at least the two most popular teacher evaluation methods now in use: value-added and supervisor evaluation scores. If teacher observations or evaluations and value-added measures are both assumed to measure the same thing, the teacher effectiveness or quality construct, then scores derived independently from both measures should represent high levels of agreement or alignment (i.e., concurrent, criterion-related evidence of validity). As well, value-added results that correlate with other respected and professionally established teacher evaluation methods would not only reduce validity concerns (see Darling-Hammond et al., 2012), but would also align with the guidelines developed by AERA, APA & NCME (1999).

Though the studies from Hill et al. (2011), Kane and Staiger (2012), and Sass and Harris (2012) did find significant correlations among various value-added scores and other subjective measures of teacher effectiveness, the relationships, VAMs used, and subjective measures all varied considerably. Hill et al. (2011) noted that some of the relationships between VAMs and other ratings of instruction also correlated with aspects of classroom composition. Kane and Staiger (2012) found that the relationship between value-added measures and observation scores increased with the inclusion of student survey data. Sass and Harris (2012) found the relationships between value-added data and subjective measures increased with multiple years of data, and also that subjective measures

such as principal observations provide information beyond value-added measures that can help in retention and tenure decision-making. However, all of the correlations listed were not inordinately high, typically no greater than $r = 0.50$ and, hence, not explaining more than 25% of the variance in value-added scores. Whether this is “good enough” has also not been discussed often, nor well enough (see also Harris, 2011; Jacob & Lefgren, 2008; Milanowski, Kimball, & White, 2004; Wilson, Hallman, Pecheone, & Moss, 2007).

Construct-related evidence measures the interpretive meaningfulness of a construct (Messick, 1975, 1980). A construct is a “nonobservable trait” that explains a behavior (Gay, 1996, p. 140). With value-added, construct-related evidence of validity represents the extent to which a standardized test measures student achievement and teacher effectiveness, both of which are considered “nonobservable traits.” This is a concern given the underlying dependence of VAMs on large-scaled standardized tests. Not only is it debatable whether student learning can be appropriately and accurately captured through up to two iterations of these tests (Braun, 2004), it may be an even greater stretch to assume teacher effectiveness is accounted for by the general tests being used to begin with (Goe, 2008). To further contextualize, content-related validity focuses on whether or not a test measures how well students understood the curriculum in a tested subject, whereas construct validity focuses on whether those tests are representative of student achievement or teacher quality or effectiveness. Obviously, the latter is a larger issue that has been debated for decades.

Finally, consequential validity examines the implications that tests have on society (AERA, APA & NCME, 1999), specifically the intended and unintended consequences that result from the interpretation and use of test data (Messick, 1989, 1995). Using value-added data to evaluate teachers, especially for high-stakes consequences, demands a thorough investigation of all possible consequences, both positive and negative, or intended and unintended respectively (Shepard, 1993, 1997), to truly evaluate the validity of the value-added system. Intended and unintended consequences will be discussed in separate, forthcoming sections.

EVAAS and validity. All of the previously mentioned validity issues apply to VAMs in general and the EVAAS model, but there are also a few other validity concerns specific to EVAAS. First, and as related to content validity, SAS recently acknowledged that test ceilings exist (A. Best, personal communication, January 21, 2012) which unfairly bias the value-added measures for teachers of high achieving and gifted students. SAS has yet to provide a resolution or explanation for how this will be remedied. Second, analyzing concurrent validity, we found evidence that some teachers who scored high on EVAAS simultaneously received low principal evaluation scores or vice versa, and at times, evidence that neither measure produced consistent or accurate measures of teacher quality (Amrein-Beardsley & Collins, 2012). Third, in terms of predictive evidence of validity, SAS (2012b) lists the ability of EVAAS to “proactively predict student success probabilities” by using historical data to predict grade level and graduation proficiencies as well as predict future college

success (p. 1). To ensure predictive accuracy, these projections need to be followed up to verify the correlations between the predicted and actual scores (Gay, 1996), yet although SAS claims to have this information, they only provide the sources that have verified their predictive methodology but not the actual data or sources of the correlations to validate the projections (Amrein-Beardsley, 2008; Sanders & Wright, 2008; SAS 2012d). In other words, the validity concerns and potential inconsistencies noted here merit a strong caution to all users of the EVAAS model, to prevent making “unsupported interpretations” as recommended by AERA, APA and NCME (1999).

Formative use. “How, and under what conditions, do policies intended to change teaching actually do so?” This quote, coming from Linda Darling-Hammond (1990, p. 341), appropriately leads into a discussion about using value-added data for formative purposes. Black and Wiliam (2004) describe formative use by teachers as when “information is actually used to adapt the teaching work to meet the learning need” (p. 22).

The reality is that growth and value-added reports themselves are simply sophisticated, albeit confusing reports with data, but they alone do not provide useful information on how to improve teacher quality (Goe, 2008; Tucker & Stronge, 2005). What teachers do with this information determines if the data lend themselves to formative use. Several researchers have expressed concern with how teachers can use value-added data to improve instructional practices (Eckert & Dabrowski, 2010; Goe, 2008; Harris, 2011; Kennedy, Peters & Thomas, 2012)

and only a few have attempted to provide instructions on how to use value-added in a meaningful way (Harris, 2011; Kennedy et al., 2012).

For example, in their “field guide” for school leaders, Kennedy et al. (2012) set out to discuss the “all too common gap between *having* value-added information and *using* value-added information” (p. xvi). The authors provided detailed examples of how to combine value-added data with achievement data to build a holistic professional development platform for districts, schools, and individual teachers. Although their recommendations and knowledge are commendable, it is unclear how many states or districts will realistically move forward with the same degree of resources (e.g., finances, time, staffing) as experienced by the districts in Ohio that were able to implement value-added in what the authors characterize as an effective manner (Kennedy, et al. 2012).

To-date, otherwise, there are no studies that examine or evidence how the formative use of value-added data is occurring. As showcased by the national overview study on growth and VAM use, not one state representative could articulate a plan for using the data for formative instructional purposes (Collins & Amrein-Beardsley, under review). Although understandably this may be individualized by districts, particularly in states with greater liberties and local control, the fact that not one state representative could provide an example of how schools and teachers were using the data for formative purposes is certainly troubling.

EVAAS and formative use. Although EVAAS provides diagnostic and customized reports, Dr. Sanders has acknowledged that student achievement and

teacher quality can only be improved by the development and implementation of strategies that lead to advancement (Sanders & Rivers, 1996). Yet most states and districts implementing the EVAAS model are not using the data in formative ways (Raudenbush, 2004). This is likely due to a lack of transparency in both the EVAAS model and the reports themselves. Although SAS claims to provide “easily understandable reporting” (SAS, 2012b, p. 1), an example of an EVAAS report is provided below (see Figure 3). Decide for yourself whether you find this report easy to understand, and if you were a teacher whether you would likely gain useful information from it to inform your teaching practices to better promote student learning.

**SAS® EVAAS® Teacher Value-Added Report for 2010
Houston Independent School District**

School:
Teacher:
Subject: TAKS/Stanford Mathematics, Grade 7

Year	Teacher NCE Gain	Tch Std Error	HISD Reference Gain	Teacher Comparison to HISD Ref Gain	Teacher Gain Index
2008	4.6	1.0	5.7	Below	-1.07
2009	4.0	1.0	6.3	Below	-2.36
2010	10.7	1.9	7.6	Above	1.62
3-Yr Avg	6.4	0.8	6.5	NDD	-0.11

Estimates are from multivariate, longitudinal analyses using all available test data for each student (up to 5 years). The analyses were completed via SAS® EVAAS® methodology and software, which is available through SAS Institute Inc. EVAAS, SAS, and other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2010 SAS Institute Inc., Cary, NC, USA. All Rights Reserved.

Interpreting the Teacher Value-Added Report

Use this report to evaluate how well a teacher facilitates student progress. The district gain is an estimate of the district's influence on student progress. The **Teacher Value-Added Report** compares each teacher's gain to the district gain. This comparison indicates how a teacher influences student progress in the given subject.

Teacher NCE Gain, Standard Error

The **Teacher NCE Gain** is a conservative estimate of a teacher's influence on students' academic progress estimated by using all students linked to a teacher who were tested on TAKS, TAKS Accommodated, or Stanford/Aprena in non-TAKS grades and subjects. It is expressed in state NCEs using 2005–2006 as the base year. The **Tch Std Error** provides the basis for establishing a confidence band around the Teacher NCE Gain value. One **Standard Error** is used in the statistical test reported under **Teacher Comparison to the HISD Ref Gain**. Note that this year's estimates of previous years' gains may have changed as a result of incorporating the most recent student data.

HISD Reference Gain

The **HISD Reference Gain** is the gain made by HISD in this subject and grade. This gain is expressed as an NCE, based on the state population in 2006. A positive gain indicates HISD made more progress than the state average and a negative gain indicates HISD made less progress than the state average.

Teacher Comparison

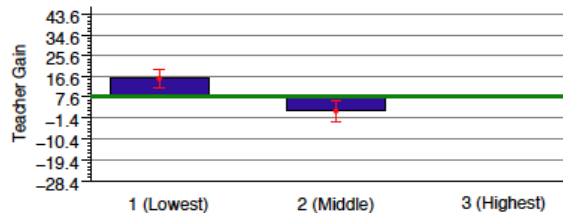
The **Teacher Comparison to HISD Ref Gain** column shows whether there is a difference in the progress rate for this teacher compared to the **HISD Reference Gain**. Comparisons are made based on one standard error.

- **Above** means that students taught by this teacher made decidedly more progress than the reference gain. A teacher classified as "above" will have Gain Index greater than 1.
- **Below** means that students taught by this teacher made decidedly less progress than the reference gain. A teacher classified as "below" will have a Gain Index less than -1.
- **NDD** means that the progress made by the teacher's students was Not Detectably Different from the reference gain. A teacher classified as "NDD" will have a Gain Index between -1 and 1.

Teacher Gain Index

To calculate the **Teacher Gain Index**, first subtract the **HISD Reference Gain** from the **Teacher NCE Gain**. Then this difference is divided by the **Tch Std Error**. The **2010 Teacher Gain Index** is used by HISD for determining ASPIRE Awards.

SAS® EVAAS® Report for Teacher Reflection, 2010



		Average Simple Gains by Prior-Achievement Subgroup		
		1 (Lowest)	2 (Middle)	3 (Highest)
Year				
2010	Ref Gain	7.6	7.6	7.6
	Avg Gain	15.7	1.6	
	Std Error	4.1	4.9	
	Nr of Students	22	9	2

Interpreting the Report for Teacher Reflection

Use this report to identify patterns or trends of progress among groups of students at different achievement levels. This report is intended for diagnostic purposes only, since students with missing scores in the current or previous year are excluded. Therefore, you will not be able to calculate Teacher NCE Gain from the table above.

The Chart

The chart offers a visual representation of average student progress for students at different entering achievement levels. The **green** line is the district reference gain line, representing the average amount of progress made by students in HISD. **Blue** bars show the progress of students in the current school year. Bars above the green line indicate that students in the subgroup made more progress than the district average. Bars below the line indicate that students made less progress than the district average. A confidence interval of one standard error is indicated in **red**. No bar is presented for subgroups with fewer than five students.

Student Assignment

Prior-achievement subgroups are determined by averaging a student's NCE score in the current year with his or her NCE score in the previous year. Students are assigned to subgroups based upon where they fall in the distribution of all students in the district who took the test in the years included in the analysis. As a result, some teachers may find that certain subgroups are more heavily populated than others.

The Table

The table immediately below the chart allows for a comparison of the subgroup NCE gains to the HISD reference gain. The standard error is reported immediately below the Average NCE Gain for each subgroup. The row labeled **Nr of Students** shows the number of students in the subgroup. When there are fewer than five students in a subgroup, the number of students in the group is reported, but the gain is not displayed.

Figure 3. A teacher's EVAAS teacher value-added report for 2010 and EVAAS report for teacher reflection.

Intended consequences and claimed benefits of EVAAS. VAMs

generate data (e.g., estimates of effectiveness) that are used to determine the impact that districts, schools, and teachers have had on student learning. EVAAS is a “comprehensive reporting package” that “provides valuable diagnostic information” (SAS, 2012a). In other words, EVAAS provides data. The simplicity of that statement is not to undermine the sophistication of the data, nevertheless, the main intended consequence of EVAAS is to provide a source of analytic data about districts, schools, and teachers. The downplayed consequence affiliated with these data is formative use by teachers, again, which Dr. Sanders acknowledged is key to improving teacher quality and made possible through the “easily understandable reporting” of the EVAAS data (Sanders & Rivers, 1996; SAS, 2012b). To determine the consequential validity of EVAAS, it is necessary to investigate the intended consequences and claimed benefits of the model.

SAS is one of the first companies to spell out, or promote, the exact benefits that schools, districts, and states can expect when utilizing their product – value added modeling. With over 20 years of development, and as previously noted, EVAAS is the largest, most widely implemented VAM in the country, and it is considered “the most comprehensive reporting package of value-added metrics available in the educational market” (SAS, 2012b, p. 1). With the ability to account for teacher and student mobility, EVAAS “provides the most reliable estimate of effectiveness of a district, school or teacher” (SAS, 2012b, p. 3). EVAAS “predict[s] student performance with precision and reliability,” “provides valuable diagnostic information about [instructional] practices,” helps educators

become more proactive and make more “sound instructional choices,” and helps teachers use “resources more strategically to ensure that every student has the chance to succeed” (SAS, 2012a). These claims and intended consequences are not without controversy, however (Amrein-Beardsley, 2008; Sanders & Wright, 2008), and I used these also to formulate questions for this study.

Unintended consequences. The consequential validity of VAMs and the EVAAS specifically, is significantly influenced by the unintended consequences that occur as a result of model implementation as well. Researchers have found little to no evidence that evaluating teachers based on student growth scores from high-stakes testing actually increases teacher quality or student achievement (Baker et al., 2010; Harris, 2011; see also Nichols, Glass & Berliner, 2006, 2012). To make matters worse, the Race to the Top competition, developed with the intentions of improving teacher quality and student academic achievement through state implementation of teacher evaluations tied to test-based growth (see USDOE, 2009), may actually be harming students (Baker et al., 2010).

As teachers succumb to pressure for their students to score highly on tests, they may either get better at training students for the tests (Nichols et al., 2012), teach to the tests, or even less desirable, cheat (Amrein-Beardsley, Berliner, & Rideau, 2010). When teachers are able to get their students to perform significantly higher on large-scaled standardized tests, it too often comes at the sacrifice of a well-rounded curriculum that encourages critical thinking, problem solving, and other skills demanded by the global knowledge economy.

Another reality, though often overlooked, is that VAMs are only able to produce value-added estimates for a minority of teachers. Again, approximately 70% of teachers do not receive value-added scores (Harris, 2011), and entire campuses (e.g., early childhood and high school buildings) can lack value-added data (Collins & Amrein-Beardsley, under review).

In addition, among those who can be evaluated by value-added, teachers who work with certain populations or grade levels are shown to be negatively impacted by bias (Carey & Manwaring, 2011; Hill et al., 2011; Newton et al., 2010). Specifically, teachers who teach special education, English language learners (ELLs), and gifted students have an unfair disadvantage when teaching these special populations, and evidence suggests these teachers are feeling disincentivized and hypothetically or actually choosing to avoid teaching these populations (Amrein-Beardsley & Collins, 2012; Hill et al., 2011).

In fact, VAMs do not appear to be robust enough to handle much separation from perfect random student assignment (Capitol Hill Briefing, 2011). Yet student sorting, or the grouping of students with similar characteristics, is a common practice in most schools. Rothstein (2010) tested the impact of student assignment by looking for value-added effects that should not exist; for example, a fifth grade teacher's effect on third grade student test scores. Through his analyses, Rothstein found that fifth grade teachers have large effects on third grade test scores – a scenario that is not even realistic. Briggs and Domingue (2011) found the same results using data released by the *LA Times*, assessing teacher effects on students they have not yet taught. These future teacher effects

were found to be similar to, and in one case even larger than those effects of the students' current teachers, indicating that student characteristics and student assignment significantly effect VAM scores, perhaps more so than teachers. Such studies indicate that value-added estimates can be seriously biased by student assignment and class composition.

These are examples of some of the unintended consequences that often arise when the underlying purpose of accountability measures is to increase student achievement, no matter what it takes. Other unintended consequences, specific to EVAAS, are discussed in further next, in Chapter 3 as related to EVAAS use in HISD, and also detailed more in the results in Chapter 5.

CHAPTER 3

The Situation in Houston

HISD is the largest school district in Texas and the seventh largest district in the country. The district consists of 300 schools, over 200,000 students, and approximately 13,000 teachers. In addition, the majority of the students in the district are from high-needs backgrounds, with 63% of students labeled at risk, 92% from racial minority backgrounds, 80% on the federal free-or-reduced lunch program, and 58% classified as ELLs, Limited English Proficiency (LEP), or bilingual. As stated earlier, no other school, district, or state uses the EVAAS for consequential decision-making more than HISD (Corcoran, 2010; Harris, 2011; Mellon, 2010; Otterman, 2010; Papay, 2010).

In 2007, HISD created the Accelerating Student Progress: Increasing Results & Expectations (ASPIRE) program, a merit-pay program developed to recognize and reward great teaching as measured by student progress (HISD, 2010). At the same time, district administrators began contracting with the SAS software company to measure this progress via their EVAAS system, at an approximate cost of \$500,000 per year.

In short, the district has two main teacher evaluation and accountability systems: 1) the ASPIRE program in which the district uses one year of EVAAS scores to rank order teachers throughout the district and 2) the Professional Development and Appraisal System (PDAS) in which teacher observation data are collected by certified appraisers (oftentimes the principals) and used to

evaluate teachers in eight different domains of teacher performance.¹³

Considering the two different foci, however, it is common that the district labels and rewards HISD teachers differently across systems, for example, labeling a teacher below average on the PDAS while rewarding the teacher with a bonus through the ASPIRE program or vice versa. The district's oft-conflicting systems cause a fair amount of confusion and mistrust, in particular among HISD teachers (Corcoran, 2010; Harris, 2011; Papay, 2010).

Before conducting this dissertation study, I participated in preliminary research in the HISD (again, along with Dr. Amrein-Beardsley) to better understand what the EVAAS looked like in practice, as well as to determine the merit for extended EVAAS research within the district. This preliminary research included focus group discussions and an examination of data collected from terminated teachers (for the full study, see Amrein-Beardsley & Collins, 2012).

Focus Group Discussions

During the spring of 2011, we conducted two focus group discussions with approximately 25 HISD teachers to discuss their experiences with EVAAS. The

¹³ During the 2010-11 academic year, HISD educators and community members helped design a new Teacher Appraisal and Development System that went into effect during the 2011-12 academic year, replacing PDAS. According to one of the district's Analysts for Accountability and Rewards, HISD plans to use student value-added data as one component of this appraisal system beginning in the 2012-13 academic year (S. Mason, personal communication, April 19, 2012).

25 teachers who participated in the focus groups responded to an open invitation to participate by the Houston Federation of Teachers (HFT), the teachers' union in Houston. There is a strong union presence in HISD, whereas approximately half of the teachers are HFT members (Z. Capo, personal communication, August 13, 2012). HFT has proactively supported their teachers' navigation through the implementation of EVAAS within the district, both morally and legally in the termination cases of teachers who were fired at least in part due to their EVAAS scores. Using my analytical framework, I organized the findings from the focus groups by issues of reliability, validity, formative use, and unintended consequences.

Reliability. Even though HISD reported that the majority of teachers favor the ASPIRE program overall (Harris, 2011), teachers participating in the focus groups indicated this was not the case, even among those teachers who have received large bonuses. Teachers who received merit monies as a result of their EVAAS output compared winning the rewards to “winning the lottery,” given the random, “chaotic,” year-to-year instabilities they have witnessed. Teachers do not seem to understand why they are rewarded, especially because they profess that they do nothing differently from year to year as their EVAAS rankings “jump around” (see also Baker et al., 2010; Corcoran, 2010; EPI, 2010; Newton et al., 2010; NRC, 2010; Papay, 2010; Rothstein 2010). Although teachers appreciated monetary awards, what they did differently from one year to the next remains unknown. For example, one eighth grade advanced English teacher noted:

I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue.

A 7th grade history teacher classified her past three years as “bonus, bonus, disaster.” Teachers who moved to different grade levels reported switching value-added ranks after the move, “flip-flopping” from “ineffective” to “effective” or vice versa, even across grade levels that were adjacent. A social studies teacher noted:

We had an 8th grade teacher, a very good teacher, the “real science guy,” [who was a] very good teacher...[but] every year he showed low EVAAS growth. My principal flipped him with the 6th grade science teacher who was getting the highest EVAAS scores on campus. Huge EVAAS scores. [And] now the 6th grade teacher [is showing] no growth [as an 8th grade teacher], but the 8th grade teacher who was sent down [to the 6th grade] is getting the biggest bonuses on campus.

This is problematic as the EVAAS system is purported to measure the teacher effectiveness construct consistently and validly, even across grade levels and subject areas (LeClaire, 2011), yet teacher mobility, especially within such a large urban district, is nothing out of the ordinary.

Validity. The majority of the validity issues discussed were content-related validity issues. Teachers who did not receive bonuses attributed the lack of monetary rewards to the types of students they taught and how these students might have biased their scores (see also Hill et al., 2011; Newton et al., 2010; Rothstein, 2009). As well, teachers reported having varied EVAAS scores despite using the same teaching techniques year-after-year, but of course with different sets of students (which is related to their concerns about bias).

Teachers who loop or teach back-to-back grade levels reported bonuses for the first year and nothing the next as they “maxed out” on growth the first year with the same students. Teachers of grades in which ELLs transitioned into mainstreamed English-only classrooms also reported being the least likely to demonstrate added value and the most likely to be deemed “ineffective.” One 4th grade teacher noted:

I went to a transition classroom, and now there’s a red flag next to my name. I guess now I’m an ineffective teacher? I keep getting letters from the district, saying ‘You’ve been recognized as an outstanding teacher’ ... this, this, and that. But now because I teach English Language Learners who ‘transition in,’ my scores drop? And I get a flag next to my name for not teaching them well?

A 5th grade teacher added:

I’m scared to teach in the 4th grade. I’m scared I might lose my job if I teach in a[n] [ELL] transition grade level, because I’m scared

my scores are going to drop, and I'm going to get fired because there's probably going to be no growth.

Another prevalent concern involved ceiling effects, whereas teachers of high performing and gifted students expressed difficulties demonstrating added value when their students consistently scored highly on tests from the start (see also Wright et al., 1997). They reported being able to “only get them up so much!” One teacher working with gifted students noted:

Every year I have the highest test scores, [and] I have fellow teachers that [sic] come up to me when they get their bonuses...One recently came up to me [and] literally cried, ‘I’m so sorry.’... I’m like, ‘Don’t be sorry...It’s not your fault.’ Here I am...with the highest test scores and I’m getting \$0 in bonuses. It makes no sense year-to-year how this works.... How do I, how do I, you know, I don’t know what to do. I don’t know how to get higher than a 100%.

Teachers of inordinate numbers of special education students expressed similar concerns demonstrating EVAAS growth (see also Hill et al., 2011; Newton et al., 2010; Rothstein, 2009). Teachers agreed that it is best for them “to get average kids, yes, because the regular kids, you can grow those kids.”

Numerous teachers, especially science and social studies teachers who taught subjects that were not tested in every grade level, noted issues and concerns when norm-referenced tests were used with criterion-referenced tests to determine EVAAS growth from year to year. Additionally, they discussed the

norm-referenced tests not being linked to state standards and the curriculum.

Although norm-referenced and criterion-referenced tests can be normed, and this is somewhat common, this still raises issues with content alignment and content-related evidence of validity.

The HISD teachers also noted that their EVAAS reports did not match their supervisors' observational PDAS scores, which pertains to criterion-related evidence of validity. Some of the teachers suggested that their principals changed their PDAS scores to reflect the findings of the EVAAS scores, given pressures from district or site administration to do so. Such adjustments provide a false representation of criterion-related validity, in that it seems the more "subjective" set of observational scores are at least in some cases being skewed to match the perceptibly more "objective" EVAAS. One social studies teacher stated:

Here's the problem: No principal wants to be called in by the superintendent or another superior and [asked], 'How come your teachers show negative growth but you have high evaluations on them? Are you doing your job? I don't understand. Your teacher shows no growth but you have [marked them] as exceeding expectations all up and down the chart?' Now it's not just this [sic] data over here that's gonna harm us, it's the principals [who are] adjusting our data over there to match the EVAAS. So it looks like they're being consistent.

An 8th grade teacher agreed adding:

They're not about to go to bat [for us, although] a few of them will. But most of them are going to go in there, and they're going to create a teacher evaluation [report] that reflects the [EVAAS] data because they don't want to have to explain, again and again, why they're giving high classroom observation assessments when the data shows [sic] that the teacher is low performing.

A 4th grade teacher noted, "Our principal pressures us. You bet she pressures [us]. If you don't make [EVAAS], then it goes against you in your PDAS. In a roundabout way she finds a way to put that against you." From these teachers' perspectives, it seems that many district administrators are more trusting of EVAAS and are skewing PDAS data to match.

Formative use. In terms of formative use, because EVAAS output is often received months after students leave their classrooms, teachers expressed that such output made little sense, particularly as the students were no longer under their instruction, and they were learning little about what they did effectively or how they might use the EVAAS data to improve their own instruction the following year (see also Eckert & Dabrowski, 2010; Harris, 2011). When asked if they had attended any professional development sessions offered by the district to learn how to interpret, understand, and use their EVAAS data, teachers either responded that they were unaware such trainings existed or that they did not find them helpful. This is problematic since EVAAS's principal claimed strength is to provide a "wealth of positive diagnostic information" for formative purposes (Sanders et al., 2009, p. 9).

Unintended consequences. Through the focus group discussions, many unintended consequences associated with the use of EVAAS in the district surfaced. As discussed under the validity and formative use sections, teachers have evidenced certain teaching scenarios that reduce their ability to score highly on the EVAAS such as looping, or teaching the same students in back-to-back grade levels. Teachers have also attributed low EVAAS scores to teaching high proportions of ELL, special education, and gifted students, and have commonly identified transition students as the most difficult students to teach and obtain high EVAAS scores. Additionally, teachers revealed that they are not using EVAAS data to inform their practices, the timing of the distribution of the data after the students have left their classrooms is not beneficial, and the trainings offered by HISD were viewed as irrelevant. These are examples of unintended consequences associated with the EVAAS use in HISD.

High stakes. Regardless, as HISD is using the EVAAS as a component of their teacher evaluations, EVAAS scores are taken into consideration for tenure and termination decisions. Shortly after the focus group discussions, a large number of HISD teachers' contracts were not renewed. As part of our continued investigation, Dr. Amrein-Beardsley and I turned our focus to the terminated teachers and their experiences with EVAAS.

Study of Terminated Teachers

In the spring of 2011 221 HISD teachers' contracts were not renewed for the 2011-12 school year (HISD, 2011). A number of these teachers' contracts were not renewed at least in part due to "a significant lack of student progress

attributable to the educator,” or “insufficient student academic growth reflected by [EVAAS] value-added scores.” HISD did not respond to our Open Records Request (submitted September 15, 2011) soliciting the actual number of unnamed teachers whose contracts were not renewed at least in part due to EVAAS scores in spring of 2011, however, so it is uncertain how many teachers were actually terminated for these reasons. What is known is that, according to one of the lead lawyers retained in these teachers’ defenses (A. Reichel, personal communication, June 8, 2011), a number of HISD teachers’ non-renewal letters cited these reasons for termination, and according to the Director of HFT, nearly 50% of non-renewed teacher contracts were at least in part due to EVAAS scores (Z. Capo, personal communication, April 6, 2012). We are also unaware of how many teachers pursued due process hearings, how many of them followed their due process hearings through to culmination, and how many were actually terminated after their due process hearings concluded; however, we were able to access and analyze the EVAAS data from four such terminated HISD teachers.

Specifically, in the spring of 2011 Dr. Amrein-Beardsley was invited by the previously mentioned lawyer to serve as an expert witness and testify on the behalves of four teachers regarding (a) the EVAAS in general, (b) whether EVAAS output for each teacher accurately evidenced that the teacher positively or negatively impacted student achievement and growth, and (c) whether the grounds and reasoning on which their contracts were not renewed were justifiable and sound.

The four teachers were females, from racial minority backgrounds, and taught at different elementary schools within HISD under different school administrators. All teachers taught core subject areas (reading, language arts, math, social studies, and science) in grades 3-7. Collectively, the four teachers averaged 11.8 years of total teaching experience and 7.5 years teaching in HISD. Out of the four teachers, only one taught the same subject or grade level for more than one consecutive year with the other three teachers switching grade levels, subjects, or both each year they were evaluated by EVAAS. As discussed earlier, this can significantly impact reliability and validity, even though EVAAS developers deny such claims (LeClaire, 2011).

But to investigate the terminations, Dr. Amrein-Beardsley collected numerous types of data from each of the four teachers, including their EVAAS Teacher Value-Added Reports (see again, Figure 4) and their PDAS evaluation scores which are also considered in the ASPIRE merit pay program. She analyzed each of the four teachers' cases individually, and then collectively, which allowed for us to compare both similarities and unique findings together with the findings from the focus group discussions.

**SAS® EVAAS® Teacher Value-Added Report for 2010
Houston Independent School District**

School:
Teacher:
Subject: TAKS/Stanford Mathematics, Grade 7

Year	Teacher NCE Gain	Tch Std Error	HISD Reference Gain	Teacher Comparison to HISD Ref Gain	Teacher Gain Index
2008	4.6	1.0	5.7	Below	-1.07
2009	4.0	1.0	6.3	Below	-2.36
2010	10.7	1.9	7.6	Above	1.62
3-Yr Avg	6.4	0.8	6.5	NDD	-0.11

Estimates are from multivariate, longitudinal analyses using all available test data for each student (up to 5 years). The analyses were completed via SAS® EVAAS® methodology and software, which is available through SAS Institute Inc. EVAAS, SAS, and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2010 SAS Institute Inc., Cary, NC, USA. All Rights Reserved.

Interpreting the Teacher Value-Added Report

Use this report to evaluate how well a teacher facilitates student progress. The district gain is an estimate of the district's influence on student progress. The **Teacher Value-Added Report** compares each teacher's gain to the district gain. This comparison indicates how a teacher influences student progress in the given subject.

Teacher NCE Gain, Standard Error

The **Teacher NCE Gain** is a conservative estimate of a teacher's influence on students' academic progress estimated by using all students linked to a teacher who were tested on TAKS, TAKS Accommodated, or Stanford/Aprenda in non-TAKS grades and subjects. It is expressed in state NCEs using 2005–2006 as the base year. The **Tch Std Error** provides the basis for establishing a confidence band around the Teacher NCE Gain value. One Standard Error is used in the statistical test reported under **Teacher Comparison to the HISD Ref Gain**. Note that this year's estimates of previous years' gains may have changed as a result of incorporating the most recent student data.

HISD Reference Gain

The **HISD Reference Gain** is the gain made by HISD in this subject and grade. This gain is expressed as an NCE, based on the state population in 2008. A positive gain indicates HISD made more progress than the state average and a negative gain indicates HISD made less progress than the state average.

Teacher Comparison

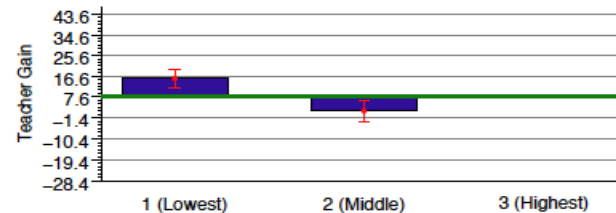
The **Teacher Comparison to HISD Ref Gain** column shows whether there is a difference in the progress rate for this teacher compared to the **HISD Reference Gain**. Comparisons are made based on one standard error.

- **Above** means that students taught by this teacher made decidedly more progress than the reference gain. A teacher classified as "above" will have a Gain Index greater than 1.
- **Below** means that students taught by this teacher made decidedly less progress than the reference gain. A teacher classified as "below" will have a Gain Index less than -1.
- **NDD** means that the progress made by the teacher's students was Not Detectably Different from the reference gain. A teacher classified as "NDD" will have a Gain Index between -1 and 1.

Teacher Gain Index

To calculate the **Teacher Gain Index**, first subtract the **HISD Reference Gain** from the **Teacher NCE Gain**. Then this difference is divided by the **Tch Std Error**. The **2010 Teacher Gain Index** is used by HISD for determining ASPIRE Awards.

SAS® EVAAS® Report for Teacher Reflection, 2010



		Average Simple Gains		
		by Prior-Achievement Subgroup		
		1 (Lowest)	2 (Middle)	3 (Highest)
Year	Ref Gain	7.6	7.6	7.6
2010	Avg Gain	15.7	1.6	
	Std Error	4.1	4.9	
	Nr of Students	22	9	2

Interpreting the Report for Teacher Reflection

Use this report to identify patterns or trends of progress among groups of students at different achievement levels. This report is intended for diagnostic purposes only, since students with missing scores in the current or previous year are excluded. **Therefore, you will not be able to calculate Teacher NCE Gain from the table above.**

The Chart

The chart offers a visual representation of average student progress for students at different entering achievement levels. The **green** line is the district reference gain line, representing the average amount of progress made by students in HISD. **Blue** bars show the progress of students in the current school year. Bars above the green line indicate that students in the subgroup made more progress than the district average. Bars below the line indicate that students made less progress than the district average. A confidence interval of one standard error is indicated in **red**. No bar is presented for subgroups with fewer than five students.

Student Assignment

Prior-achievement subgroups are determined by averaging a student's NCE score in the current year with his or her NCE score in the previous year. Students are assigned to subgroups based upon where they fall in the distribution of all students in the district who took the test in the years included in the analysis. As a result, some teachers may find that certain subgroups are more heavily populated than others.

The Table

The table immediately below the chart allows for a comparison of the subgroup NCE gains to the HISD reference gain. The standard error is reported immediately below the Average NCE Gain for each subgroup. The row labeled **Nr of Students** shows the number of students in the subgroup. When there are fewer than five students in a subgroup, the number of students in the group is reported, but the gain is not displayed.

Figure 4. Teacher C's EVAAS teacher value-added report for 2010 and EVAAS report for teacher reflection.

Here again, I applied my analytical framework to organize the findings from this study along issues of reliability and validity for each of the four teachers. After, I discuss reliability and validity, overall, along with formative uses and unintended consequences as they pertain to these four and other HISD teachers.

Teacher A. In looking at her four years of data, Teacher A added value to her students' learning (relative to all other HISD teachers) 50% of the time (8/16 of EVAAS observations), and detracted value (relative to all other HISD teachers) the other 50% of the time (8/16 of EVAAS observations; see Table 1).

Reliability. According to these EVAAS output, the probability that Teacher A was truly an effective or ineffective teacher was no different than the flip of a coin. Additionally, looking at Teacher A's most recent years of activity, she added more value than she had in previous years, making termination unreasonable and indefensible, especially on the grounds that there was "a significant lack of student progress attributable to the educator" or "insufficient student academic growth reflected by [EVAAS] value-added scores."

Table 1

Teacher A's EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010)

	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011
	Grade 5	Grade 4	Grade 3	Grade 3	Grade 5
Math	-2.03	+0.68*	+0.16*	+3.46	n/a
Reading	-1.15	-0.96*	+2.03	+1.81	n/a
Language Arts	+1.12	-0.49*	-1.77	-0.20*	n/a
Science	+2.37	-3.45	n/a	n/a	n/a
Social Studies	+0.91*	-2.39	n/a	n/a	n/a
PDAS: % of Total	98.0%	98.4%	98.4%	89.0%	53.7%
ASPIRE Bonus	\$3,400	\$700	\$3,700	\$0	n/a

Note. Scores shaded as green indicate that the teacher added value according to EVAAS data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (a) Scores with asterisks (*) did not signify statistical significance, but the opposite. They signify that the scores were not detectibly different (NDD). This means that the progress Teacher A's class made was not detectively different from the reference.

Validity. Analyzing Teacher A's EVAAS scores alongside her PDAS scores, it was visually obvious and statistically evident that there was something peculiar about the relationship between Teacher A's performance on the EVAAS

and her supervisor evaluation scores. The correlation between Teacher A's EVAAS and PDAS scores across reading ($r = -0.51$), math ($r = -0.83$), and language arts ($r = -0.11$) from 2007-2010 suggest that beyond no correlation, the better Teacher A did on the EVAAS the worse she did in the eyes of her supervisor(s), and vice versa. In addition, Teacher A was monetarily rewarded in a way that did not make sense. The worse she did the more money she received for ASPIRE ($r = -0.42$).

Teacher B. Teacher B had three years of EVAAS data; two years of negative value-added scores for math, and positive for the most recent year for which she had data (see Table 2).

Reliability. Teacher B was the only teacher out of the four who taught the same subject and grade level for more than two years. In her most recent year with data, she seemed to have added value to her students' learning. Given this positive EVAAS score, whether she demonstrated "a significant lack of student progress attributable to the educator," or "insufficient student academic growth reflected by [EVAAS] value-added scores" was debatable.

Table 2

Teacher B's EVAAS and PDAS Scores and ASPIRE Bonuses (2008-2010)

	2007-2008	2008-2009	2009-2010	2010-2011
	Grade 7	Grade 7	Grade 7	Grade 9 & 10
Math	-1.07	-2.36	+1.62	n/a
PDAS:% of Total	58.0%	55.3%	59.2%	n/a
ASPIRE Bonus	\$1,750	\$0	\$4,700	n/a

Note. Scores shaded as green indicate that the teacher added value according to EVAAS data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (a) Scores with asterisks (*) did not signify statistical significance, but the opposite. They signify that the scores were not detectibly different (NDD). This means that the progress Teacher B's class made was not detectively different from the reference gain scores of other teachers across HISD given one standard error; however, the scores were still reported to both the teachers and their supervisors as they are presented here.

Validity. Analyzing Teacher B's EVAAS scores alongside her PDAS scores, there is a strong relationship between Teacher B's EVAAS and supervisor evaluation scores ($r = 0.91$). The better Teacher B did on the EVAAS the better she did in the eyes of her supervisor(s), and vice versa. This is the type of correlation coefficient we would expect to see if both indicators reliably and validly measured teacher effectiveness (i.e., criterion-related evidence of validity). In addition, Teacher B's ASPIRE bonuses were rewarded in a way that made sense; the better she did the more money she received ($r = 0.93$).

Teacher C. Looking at Teacher C's four years of EVAAS data, she detracted from her students' learning (relative to all other HISD teachers) 100% of the time across three subject areas during her three years of EVAAS scores (see Table 3).

Reliability. For three years, Teacher C was evaluated in 6th grade social studies, with two years also including math (see Table 3). For the last year Teacher C was evaluated only in 6th grade science, which she had not taught in the past. Although she does have more than two years of social studies evaluations, her scores for 2007-08 were not detectably different from the average teacher. These reliability issues make it difficult to say whether or not Teacher C was responsible for a significant lack of student growth.

Table 3

Teacher C's EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010)

	2006-2007	2007-2008	2008-2009	2009-2010
	Grade 6	Grade 6	Grade 6	Grade 6
Math	-1.67	-2.58	n/a	n/a
Science	n/a	n/a	n/a	-1.09
Social Studies	-1.72	-0.16*	-1.14	n/a
PDAS: % of Total	84.6%	86.3%	88.6%	78.0%
ASPIRE Bonus	\$1,000	\$100	\$475	\$1,225

Note. Scores shaded as green indicate that the teacher added value according to EVAAS data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (a) Scores with asterisks (*) did not signify statistical significance, but the opposite. They signify that the scores were not detectibly different (NDD). This means that the progress Teacher C's class made was not detectively different from the reference gain scores of other teachers across HISD given one standard error; however, the scores were still reported to both the teachers and their supervisors as they are presented here.

Validity. Looking beyond the data, it was uncovered that Teacher C taught some of the highest needs students, possibly across the district. The ages of the 6th grade students in her remedial classes ranged from 10 (the typical age of a 6th grader) to 15 (the typical age of a high school freshman). Almost half of Teacher C's students over time were retained in grade one to four times prior.

Analyzing Teacher C's EVAAS scores alongside her math PDAS scores was not possible as only two EVAAS scores were available, although her social studies EVAAS and PDAS scores were mildly related ($r = 0.26$). Teacher C's ASPIRE bonuses and PDAS scores were also mildly related ($r = 0.29$).

Teacher D. In analyzing Teacher D's four years of EVAAS data, it is evident she switched back and forth across grade levels and subject areas, demonstrating added value overall from 2006-2009 50% of the time (3/6 EVAAS observations) and demonstrating negative value 50% of the time (3/6 EVAAS observations; see Table 4).

Reliability. According to her EVAAS output, the probability that Teacher D was an effective teacher up until 2009-2010 was no different than the flip of a coin. Given Teacher D's most recent year of EVAAS data (2009-2010), however, she seemingly detracted from student learning across all three subject areas.

Table 4

Teacher D's EVAAS and PDAS Scores and ASPIRE Bonuses (2007-2010)

	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011
	Grade 4	Grade 3	Grade 3	Grade 4	Grade 3
Reading	+0.36*	-0.17*	-2.28	-3.88	n/a
Language Arts	-1.60	+1.28	+0.39*	-3.25	n/a
Social Studies	n/a	n/a	n/a	-2.36	n/a
PDAS: % of Total	65.5%	71.4%	74.5%	61.6%	43.5%
ASPIRE Bonus	\$1,500	\$2,900	\$2,150	\$1,250	n/a

Note. Scores shaded as green indicate that the teacher added value according to EVAAS data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (a) Scores with asterisks (*) did not signify statistical significance, but the opposite. They signify that the scores were not detectibly different (NDD). This means that the progress Teacher D's class made was not detectibly different from the reference gain scores of other teachers across HISD given one standard error; however, the scores were still reported to both the teachers and their supervisors as they are presented here.

Validity. In 2009-2010 Teacher D was assigned to teach an inordinate number of ELLs who were transitioned into her classroom, which can make it a challenge to demonstrate evidence of high growth (Carey & Manwaring, 2011; Hill et al., 2011; Newton et al., 2010). Considering this and the lack of consistent EVAAS data, whether Teacher D demonstrated “a significant lack of student progress attributable to the educator,” or “insufficient student academic growth reflected by value-added scores” is still disputable, however.

Analyzing Teacher D's reading EVAAS scores alongside her PDAS scores, there was a mild correlation ($r = 0.29$). In terms of her performance on the PDAS and her students' EVAAS scores in language arts, there was a strong correlation ($r = 0.92$). In addition, the better Teacher D scored on the EVAAS the more ASPIRE money she received ($r = 0.79$).

Overall Findings

Reliability. Again, EVAAS is meant to “assess and predict student performance with precision and reliability” (SAS, 2012a). In terms of the data presented by the four teachers (as well as the focus group discussions) however, it is clear that data inconsistencies were a consistent problem. Yet these four teachers were removed from their teaching positions “at least in part” due to EVAAS data that in three of the four cases we evidenced as unreliable (see again, Tables 1-4). The probability that three of the four teachers added or detracted value from year to year was roughly the same as the flip of a coin. This is pragmatically, methodologically, conceptually, and morally concerning. As previously mentioned, researchers suggest that two or more years of value-added data are needed to make such judgments (AACC, 2011; Brophy, 1973; Harris, 2011; Koedel & Betts, 2009). This is also troublesome as not one of the four teachers had more than two years of consistent data (that were detectibly different from other similar teachers) to warrant non-renewal.

Validity. The EVAAS and PDAS scores for the four terminated teachers were anything but consistent. Some years teachers scored highly on the EVAAS but received low PDAS scores or vice versa, indicating that both measures lacked

concurrent validity. In addition, three out of the four teachers had received teacher of the month or year awards during the same years that their EVAAS scores at least partially contributed to their termination. Criterion-related validity implies that the two measures, assumed to demonstrate teacher quality, would produce similar results with strong correlation coefficients, with teacher awards serving as additional validation.

Formative use. Each of the four terminated teachers was asked about their use and interaction with their EVAAS reports. All four of the teachers indicated they were only familiar with EVAAS, mainly when their score reports were distributed each year or in hearing other teachers talk about “value-added.” But none of the teachers indicated having conversations with principals or administrators who were able to explain their scores or how they might use them to inform teaching practices. Teacher A, B, and C did not receive training to understand, or professional development to improve their value-added scores. Additionally, they didn’t understand what the scores meant or how to use them to inform instructional practices. Teacher D used HISD’s online training to try to understand the EVAAS but said she still found it very confusing and she was not sure how to apply the information to her teaching practices. These teachers’ statements contradict HISD’s claim that, “Teachers with poor evaluations are offered support and professional development opportunities before decisions are made on contract renewal” (HISD, 2010).

Unintended consequences. As was similarly evidenced with the HISD teachers in the focus groups, the four terminated teachers also felt pressure from

administrators regarding their EVAAS scores. The four teachers felt they were targeted for termination because of the performance of the schools in which they taught, all of which were labeled “in-need-of-improvement” following the state’s interpretation of the categories mandated by NCLB. According to the teachers, administrators were under intense district and state pressure, and set out or were forced to “restructure the school” and “start firing teachers.” Teachers A, B, C, and D all felt that they were part of “a larger plan” and had been put “on a list.” When their PDAS observational scores plummeted, the four teachers began to feel vulnerable as it seemed the district was manufacturing their cases against them.

Teacher A “exceeded expectations” on her yearly PDAS reports until 2010-2011 when a new principal arrived and ranked her “proficient” or “below expectations” across domains. Teacher B’s PDAS scores dropped as well, but her supervisor wrote on her PDAS form that she could not have earned higher scores because the state classified the school’s scores as “unacceptable.” Three different administrators evaluated Teacher C and she consistently “exceeded expectations,” but in 2010-2011 when evaluated by a short-term administrator, she too was rated as “proficient” or “below expectations” across the board. Similarly, Teacher D’s supervisor’s actions became perceptibly more aggressive. According to these teachers, it seemed that EVAAS was used as a tool to eliminate teachers rather than help them to improve their practices.

Final verdicts. Ultimately, Teachers A, B, and D pursued due process hearings, but they decided not to follow their hearings through to culmination. They decided to quit teaching in HISD or altogether. Teacher C (the teacher who

according to her EVAAS output had the poorest visible value-added scores) took her case through her due process hearing. Her hearing officer noted that the types of students Teacher C typically taught most likely biased her capacity to demonstrate value-added and show growth. The hearing officer also noted that Teacher C did not have multiple years of consistent data in the core subject areas she taught to warrant a decision regarding whether she was indeed an effective teacher. Teacher C was given her job back.

Merit for Dissertation Study

Based on the cases of these four teachers and the aforementioned results from the initial focus groups, it seemed evident that the district was and is continuing to inappropriately use inconsistent data within and across subject areas to make high-stakes decisions about teachers, including decisions about teacher termination. Even though both of these research studies (the first with the focus groups and the second with the four terminated teachers) would not generalize beyond the district or perhaps even within it given various limitations per study (see Amrein-Beardsley & Collins, 2012 for an in-depth discussion of these limitations), both of the study's findings provided much new information on what EVAAS looked like in practice. From this work, it was evident that the use of EVAAS within HISD, especially for such high-stakes consequences, merited further investigation.

CHAPTER 4

Methods

In this chapter I discuss the methodology I selected for this study. It is important to note that this study, particularly given its controversial nature, was reviewed and approved by Arizona State University's Office of Research Integrity and Assurance (see Appendix A). Additionally, I received approval from HISD's Department of Research and Accountability to conduct this study with their teachers (see Appendix B).

As detailed in Chapter 3, the preliminary research we completed in HISD evidenced there was much more to be learned about how EVAAS implementation and its high-stakes use throughout the district were impacting HISD teachers. Again, the purpose of this study was to gain an understanding of how teachers and their teaching practices have been impacted by the implementation and use of the EVAAS model to hold them accountable within HISD. The primary research question for this study was: What are the intended and unintended consequences, as experienced by HISD teachers, through the implementation and use of the EVAAS model?

Survey Research Study

Survey research studies allow for the simultaneous "examination of hundreds or even thousands of survey respondents" (Babbie, 1990, p. 41). In using a survey, a researcher can infer information about a population based on responses from a set of sampled participants, sampled from the population to which results are intended to generalize (Gay, 1996). Although surveys intended

to study attitudes or opinions are more difficult to design and require careful definition of the population to be sampled, the descriptive and explanatory data yielded can provide very valuable information (Gay, 1996), and in this case the information desired given the purpose of this study. The survey method also allows for interplay between qualitative and quantitative measures (Strauss & Corbin, 1998). This offers more than simply using qualitative data to support quantitative findings or vice versa, as it actually helps to capture participants' beliefs and actions (Strauss & Corbin, 1998).

In this case, conducting a survey research study allowed me to ask HISD teachers questions about both their teaching practices and their perspectives about the EVAAS and its use within HISD. Given the large quantity of teachers in HISD, a survey research method was also most appropriate to engage as many teachers as possible, and to better and more comprehensively learn about their perspectives and experiences with the EVAAS model as a whole. As surveys can be designed and distributed electronically, I was able to investigate the largest potential number of teachers in Houston, also while remaining in Phoenix.

Additionally, this format allowed me to carefully design specific constructs within the survey to correspond with my analytical framework. Again, the constructs specifically included: reliability (whether teachers received consistent EVAAS scores from one year to the next); validity (whether EVAAS scores matched other measures or evidence of teacher effectiveness); formative use (whether teachers used EVAAS data to inform their instruction); and intended consequences and claimed benefits of EVAAS (whether teachers agree with

EVAAS marketing statements and claimed benefits). Acknowledging the difficulties associated with attitudinal research designs (Gay, 1996), it is my main intention to describe both my mixed methods approach as precisely and thoroughly as possible to demonstrate trustworthiness in the ultimate findings (Babbie, 1990).

Mixed Methods Approach

“A mixed methods approach is one in which the researcher tends to base knowledge claims on pragmatic grounds” (Creswell, 2003, p. 18). The researcher collects both numeric information and qualitative texts to produce a mixed database including quantitative and qualitative information (Creswell, 2003). The mixed methods approach allows researchers to combine the strengths, while minimizing the weaknesses, of both quantitative and qualitative approaches in one single study, while also providing an optimal opportunity to investigate research questions (Johnson & Onwuegbuzie, 2004). This optimal combination of mixed methods allows for more insight and more comprehensive findings (Greene, Benjamin & Goodyear, 2001).

I selected a mixed methods approach given, again by its nature, it would help me to investigate the realities of EVAAS from the perspectives of teachers, by quantifying both common and individual findings, while also providing contextual support for these numerical responses through the actual words and stories of the teacher participants. Additionally, the quantitative and qualitative findings from this study could be compared, or triangulated across data sources and along with the literature on this topic (Bernard & Ryan, 2010). The mixed

methods approach also allowed me to examine the quantitative and qualitative data generated from the large-scale survey research study at one time.

Further, and as explained by Creswell (2003), mixed methods allow for greater priority to be given to either the quantitative or qualitative approach (p. 212). This preference depends on the researcher and also the emphasis of the study. With the goal of this study to learn from the experiences and perspectives of teachers, I chose the qualitative approach to be more dominant as it was the teachers' own words and stories in which I was most interested for this study. Nonetheless, quantifying some of their responses, particularly for descriptive purposes was also a priority.

Quantitative data. In a mixed methods approach, quantitative inquiries such as Likert-type scales can be used to complement forms of qualitative inquiry (Morse, 2003). A Likert-type scale is a psychometric instrument used to ask a series of questions that are designed to measure attitudes, opinions, and beliefs (Coladarci, Cobb, Minium & Clarke, 2008). Likert-type scales are typically ordinal, meaning that the statements can be interpreted as agreeing more or less, but precisely how much more someone agrees with a statement cannot be measured (as is the case with continuous measures).

In this case, I used a Likert-type scale to quantify some of the greater attitudes and beliefs of HISD teachers about the EVAAS and its implementation in HISD. Specifically, I constructed a Likert-type scale that contained a series of 16 statements meant to yield teacher participants' beliefs about EVAAS marketing statements, claimed benefits, and overall statements about the EVAAS

and its implementation. I assigned the following point values to responses accordingly: Strongly Agree = 5, Agree = 4, Neither Agree nor Disagree = 3, Disagree = 2, Strongly Disagree = 1 (Gay, 1996, p. 155).

Qualitative data. Mixed methods allow for qualitative findings and results that portray the stories and lived experiences of a set of participants as well (Greene, 2008, p. 7), and the qualitative findings and results allow meaning to be made from investigating participants' lived experiences and perspectives (Creswell, 2003; Strauss & Corbin, 1998). This, in particular, allows for substantive exploration in areas in which little is known (Strauss & Corbin, 1998).

Analyzing qualitative texts also allows researchers to find and code themes, which are typically manifested through repetition in responses (Bernard & Ryan, 2010). The "constant comparison method," a method that will be discussed more momentarily, allows researchers to compare and contrast across the texts from various subjects (Glaser & Strauss, 1967), as well as various forms of data. Strauss and Corbin explain (1998) if qualitative data are analyzed and portrayed correctly, "then we are not speaking for our participants but rather are enabling them to speak in the voices that are clearly understood and representative" (p. 56). This was my intent here, and the main reason my priority in designing this study was to gather data to allow the voices of HISD teachers to be heard. Through careful qualitative data collection and analyses, my ultimate goal was to be able to accurately portray the experiences and perceptions of HISD teachers, allowing their voices to be heard and understood.

As well, as most value-added studies conducted to date have been heavily quantitative, the qualitative analysis of teachers' experiences with the EVAAS model provides insight into the realities of the model as it is used in practice (i.e., from theory to practice). As such, this is the first study to illuminate the realities of at least this particular value-added system in high-stakes use, and this is the first response to other topical researchers calling for such studies, mainly to examine "the relationship between value-added scores and the characteristics they are assumed to represent: good teaching, and by extension, good teachers" (Hill et al., 2011, p. 795).

Survey Instrument

In constructing the survey instrument, I first listed all of specific topics I wanted to turn into questions that fell within each of the four constructs included in my analytical framework (Babbie, 1990). Second, I created a list of EVAAS statements from the SAS website and literature and used these to develop the Likert-type scale to which I referred before, for teachers to indicate their level of agreement with each of the statements. Though quantitative results from Likert-type items can make analyses less messy (Babbie, 1990), it was also important that teachers did not feel limited in their ability to respond, and that they felt free provide more unique and personal information about the impact EVAAS has had on them and their teaching practices. So third, I created both open- and closed-ended questions to best capture teachers' individual experiences, beliefs, and opinions. Related, each of the closed-ended questions (i.e., Likert-type items) encouraged respondents to expand further should they have so chosen. Each of

the qualitative open-ended questions encouraged teachers to explain their own personal experiences and perceptions, again about the EVAAS and its use in HISD.

In total, the final survey instrument contained 12 demographic items, four reliability items, five validity items, nine items regarding formative uses and summative consequences, seven open-ended items regarding reliability, validity, formative use, how to obtain the highest EVAAS score, and overall open feedback, 11 items requesting further explanations if desired, and 16 Likert-type items related to EVAAS's intended uses, claimed benefits and overall opinions of EVAAS. None of the items required an answer, which resulted in participants skipping some questions (see the complete Survey Instrument in Appendix C). I did not impute values because the questions that respondents skipped were skipped at random, and as such would not result in uncertainty as the majority of respondents answered all questions. Additionally, I did not want to add additional bias by imputing values for missing data based on respondent means, particularly given the strong union presence in the sample (discussion forthcoming in Chapter 5).

For content validation and clarification purposes during the instrument design process, I received feedback from Dr. Amrein-Beardsley and Zeph Capo, the Director of HFT (referred to in Chapter 3 and who helped organize the previously discussed focus groups), and I also met in-person with a group of HISD teachers to ensure I captured the main issues with EVAAS. Additionally, I invited a separate group of seven HISD teachers who had previously participated

in the focus group discussions (also referenced in Chapter 3), to pilot the survey online to test for the reliability of the instrument as well as general clarity.

I determined clarity by informally assessing teachers' understanding of the questions, and whether their responses were at least close to what I intended to capture. I determined whether sufficient levels of reliability were posted through the calculation of Cronbach's alpha, which measures whether the questions were asked in a consistent, related manner, and that questions did not solicit conflicting responses (Cronbach, 1951). In general, the higher the alpha value the more reliable the test, where $\alpha \geq 0.70$ is considered adequate (Nunnally, 1978). Cronbach's alpha for the pilot instrument was $\alpha = 0.92$ and the alpha for the final instrument was $\alpha = 0.96$. In other words, the questions on the instrument yielded very consistent data across respondents. Piloting the survey instrument also ensured there were no glitches in the survey software program, and also allowed me to keep track of how long it took pilot participants to complete the survey.

I used Remark Web Survey Software developed by Gravic (Gravic, 2012) to administer the survey instrument to all participants. The software offers sophisticated and customizable features, such as custom URL and survey design features. Additionally, the survey was hosted on Arizona State University's server, which removed concerns about data storage capacity and security. I was able to work with both Remark and ASU staff for technical assistance.

Participants

As per HISD approval to conduct research, "the [researcher is] responsible for identification of the survey population." The number of EVAAS eligible

teachers (i.e. those who received individual EVAAS scores) for the 2011-12 ASPIRE bonus awards was 4,397 (L. Zimmerman, personal communication, March 8, 2012), approximately 38.8% of all teachers in HISD (including charter school teachers). This is similar to what other researchers have noted about how many teachers are included in these teacher accountability systems based on VAM output (i.e., approximately 30%; see for example Harris, 2011); however, I did not just contact these 4,397 because I did not want to exclude any teachers who taught in an EVAAS eligible subject area or grade level prior (e.g., K-2 teachers who might have taught in other elementary grade levels in years prior).

I received valid email addresses for 6,292 K-8 teachers from an HFT employee who verified the list with HISD, all the while understanding that many of the 6,292 teachers may not have been EVAAS eligible, and they might not have ever received EVAAS data. Ideally, I would have contacted all of the HISD teachers who had been employed by HISD for the past five years and were EVAAS eligible. However, as per the HISD approval to conduct research, I was to identify and collect the email information for the sample I was to draw from the HISD population. Without district assistance, I did not have the means to identify the exact number of HISD teachers who had ever received EVAAS scores, which made it difficult to narrow in on a precise sample, or related, an exact denominator.

Regardless, I distributed the survey instrument via email to all 6,292 K-8 teachers on February 3, 2012, with an explanation of the study and an invitation to participate in the voluntary study *if* the HISD teachers solicited were indeed

EVAAS eligible (see the email I sent to these teachers in Appendix D). I kept the survey link live through the month of February, and the same 6,292 teachers were sent reminder emails on a weekly basis. I personally emailed potential teacher participants three times including my initial email and two reminder emails (see also Appendix D), and Dr. Amrein-Beardsley emailed all potential participants an email reminder once (see also Appendix D). I closed the survey on March 5, 2012, as well as data collection. I was able to export all data from Remark into Microsoft Excel from there to begin data analyses.

Unbeknownst to me, however, was that HFT sent a reminder email to all HISD employees prior to the end of data collection (L. Zimmerman, personal communication, March 8, 2012). For this reason, it is also unclear how many teachers (beyond the 6,292 that I had contacted) received an email invitation to participate, and therefore, again, it is even more impossible to determine a precise denominator. This also required me to hand clean the data to ensure I included only the data for EVAAS eligible teachers in the analyses.

Data cleaning. When the survey closed on March 5, 2012, the total number of teachers who had responded to the survey was 1,338. Off the bat, I reduced the 1,338 respondents to 1,323 as 15 teachers either indicated they were not currently employed by HISD or did not answer that question. I then eliminated 284 teachers who did not respond that they had ever taught grade levels 3-8, which reduced the sample size to 1,039. Then I removed 133 teachers who responded they had never received an individual EVAAS report, reducing the sample size to 906. Finally, through reading each survey response line-by-line,

I removed the responses for 24 additional teachers who indicated via open-ended questions that they were physical education, art, music, life skills, test preparation, or pre-k teachers, and therefore ineligible. This careful data cleaning reduced the sample size to 882 teachers.

Response rate. Again, as the total number of teachers who were emailed the survey via HFT was unknown, and as the email list I initially used included a fair amount of HISD teachers who were EVAAS ineligible, I calculated the response-rate using the number of teachers included on my email distribution list, 6,292 (882/6,292; 14.0%), although this is likely an underestimate given the aforementioned issues. Had I used the 4,397 teachers who were eligible for EVAAS data in 2011-2012, asserting that each year approximately the same number of teachers were eligible per year (around 38.8%), the response rate would have been 20.1% (882/4,397), although this would have likely been an overestimate instead. That said, I estimate that the actual response rate is probably within the range of 14.0%-20.1%.

Further, I used a confidence interval calculator with a 95.0% confidence level to determine that the sample size needed to support generalization was indeed achieved for the more conservative 14.0% response rate (Creative Research Systems, n.d.). The margin of error for the total sample is +/- 3.06% at the 95.0% confidence level. As such, I will refer to the response-rate as 14.0% hereafter (882/6,292).

There were many factors that may have impeded a higher response rate and they include the following. First, although this study was approved by the

district, it was not endorsed by the district. I was solely responsible for distributing the survey via email. It is possible that teachers overlooked or deleted my emails since they did not recognize my email address. Second, teachers have little time to check their email during the school day and are already inundated by other electronic surveys sent to them by the district. Additionally, HISD administrators made it very clear that the survey was not to be completed during instructional time. Finally, given that EVAAS is a high-stakes topic in the district, teachers may have been wary that despite my promises of anonymity, their identities could have been in some way discovered and their responses possibly used against them.

Generalizability. The low response-rate also reduces the ability to generalize the findings of this study beyond Houston and this sample of teachers. Although I determined that the sample size needed to support generalization was achieved (Creative Research Systems, n.d.), it could still logically be argued that only the most vocal or opinionated teachers participated in this study, which still makes it unclear whether the results of the study can be generalized to other HISD teachers who are also impacted by EVAAS.

Creswell (2003) described generalizability as “the external validity of applying results to new settings, people, or samples” (p. 195). Qualitative studies, or in this case a mixed method study with a lot of qualitative data, provide a lot of data about a limited number of cases, or even a single case which does not lend to generalizing results to a larger population (Dey, 1993). Given HISD’s unique use of EVAAS within the ASPIRE program, and the high-stakes consequences

attached to the EVAAS output (termination and merit-pay), the generalizability of the results of this study is still questionable.

However, it is still possible to learn even from “internal generalizations,” (Maxwell, 1941) where the findings from the HISD teachers in this study show trends within one particular setting or group studied that seem similar to each other. In this case, there were many similarities among findings and among the sampled teachers, again, which included both union and non-union teachers. Additionally, by applying Stake’s “naturalistic generalizations,” the readers might better gain insight from my description of the EVAAS situation in HISD through the words and experiences of teachers. Here, it becomes the readers’ responsibility to generalize from the findings within their own contexts and given their own experiences and constructed realities (Stake & Trumbull, 1982).

Data Analyses

I analyzed the data resulting from the survey responses from the 882 teacher participants included in the final dataset. For the quantitative analyses, I uploaded the survey responses into IBM SPSS statistical software for statistical analyses. First, I calculated means and standard deviations for each question. Then I calculated other descriptive statistics, including response rates per item and for all questions (Gay, 1996).

Through initial analyses, I also realized that approximately two-thirds (69.4%) of the respondents were members of HFT. Accordingly, I calculated chi-square analyses for each statement which examined whether the perceptions among the two mutually exclusive groups, union and non-union teachers were

significantly different (Gay, 1996). Chi-square analyses help to examine whether a person's response is contingent upon a certain characteristic (Coladarci et al., 2008), which in this case represented union or non-union status. I present all of my findings from these analyses in Chapter 5, but for now it is important to note that all analyses illustrated that there was no statistically significant difference between the responses of HFT members and non-HFT members. In other words, non-union and union members had essentially the same thoughts and beliefs about the EVAAS system and its use within the district. Being a member of the union did not significantly bias respondents' one way or another per issue of interest.

Otherwise, the open-ended, free-response responses of the 882 teacher participants ultimately yielded 4,594 unique responses in total. I analyzed these qualitative data in Microsoft Excel and printed hard copies to complete coding by hand (Miles & Huberman, 1994). Specifically, when coding I focused on one survey construct at a time, moving through the constructs the same way in which they were presented in the survey protocol as aligned with my analytical framework. I analyzed the responses line-by-line (Strauss & Corbin, 1998). While doing this, I also used open coding where codes were assigned to each response based on what made the most sense (Strauss & Corbin, 1998), keeping track of the number of respondents for each code. I did not employ preconceived codes, but I let the data speak for themselves (Miles & Huberman, 1994; Straus, 1987). Related, I also discovered unintended consequences as a result of analyzing the data across constructs along the way.

Repetition was fundamental in creating my first round of codes (Strauss, 1992) as was Glaser & Strauss's (1967) "constant comparison method" to look for similarities and differences across the data. This also helped me to ensure that I did not force a response into an existing code (Glaser, 1978; Strauss & Corbin, 1990). After the initial round of coding, I had created 10-12 codes for each open-ended question, which included the proportion of respondents for each code.

Next, I applied Lincoln & Guba's (1985) "cutting and sorting" technique where I collapsed the codes from each individual question into larger subgroups of codes, again keeping in mind the number of respondents for each code. Once again, I used constant comparison to look for similarities and differences within coded subgroups (Dey, 1993) and created both larger and smaller groups of codes that were "more conceptually inclusive" and represented "more differentiated instances" respectively (Miles & Huberman, 1994). I used the remaining subgroups to create the main themes (Lincoln & Guba, 1985). From this process, I identified three main themes for each of the survey constructs. Finally, I collapsed and quantified all qualitative themes to numerically represent them (Miles & Huberman, 1994), and I ultimately used these data along with their frequencies when triangulating the findings.

Validity

Triangulation is the use of quantitative data and literature, or additional qualitative data to support the qualitative findings within a study (Creswell, 2003), serving as a cross-check among data sources or verification of the findings (Gay, 1996). For this study, the survey data served as the primary data source and

several other forms of secondary data were used to triangulate my findings and understandings. I collected and researched as many documents as I could find on both the EVAAS and ASPIRE program via the HISD website. I had one 45-minute phone conversation with Carla Stevens, HISD's Assistant Superintendent of Research and Accountability to clarify and verify the information I had collected about the EVAAS and ASPIRE program, as well to help me in my understanding of the modifications made to determine ASPIRE rankings over time. I exchanged approximately 10 emails with HISD employees in the Department of Research and Accountability, and approximately 5 emails with Zeph Capo from HFT to also authenticate information along the way.

Although I did not conduct the same extensive analyses on these secondary sources of information as I did the survey data, the documents, phone conversations, and email exchanges all served very important purposes in terms of supporting the overall validity of this study; that is, verifying information I found online, clarifying misunderstandings, and also when identifying the teachers eligible to participate in the EVAAS survey research study were all important and methodical processes.

To enhance the validity of the inferences I was to draw via the survey data, line-by-line analyses of all of the qualitative data, again, ensured all voices and perspectives were heard and kept intact. I used teacher quotations as often as possible in presenting the findings to demonstrate participants' lived experiences and to ensure their voices are heard throughout the results (Creswell, 2003; Strauss & Corbin, 1998). This also enhanced validity.

Additionally, I completed member checks (Lincoln & Guba, 1985), which allowed a group of five, self-selected teacher participants to read and respond to the main themes as derived via data analyses, as well as the overall findings of the study. Teachers indicated that the themes and overall findings were consistent with what they had experienced and what they had heard from other teachers. One teacher said, “The study highlighted that SAS EVAAS is full of holes and yet the district relies on it. How can this be?” Another teacher said “I think you nailed it. It is appropriate to ask for teacher input regarding this practice. So often, administrators and school board members are asked their opinions with little or no input from those most directly affected.” Another said, “The results are aligned with what we have seen as teachers and what we have experienced throughout this ‘grand’ experiment on our kids and teachers.” This process of member checking ensured my findings were representative and on target.

Role of the Researcher

As I took careful measure in all of these efforts to increase the trustworthiness of my findings, I also took careful measure to reduce my personal bias as the primary researcher in this study. Ultimately the reader will determine my credibility, but I believe I presented the data in as honest and pure form as possible, and that the voices of teachers will speak for themselves, and ultimately speak more truth to power than I am able to provide.

Namely, my role as a researcher in this study was a non-participant, researcher; however, I entered into this study with a high-level of knowledge about value-added research. I was also fully aware of the high-stakes situation in

HISD (as evidenced in Chapter 3), which is why I selected this district as the site for my preliminary and further investigations. Notably and throughout the duration of the study, I co-authored published articles evidencing my concerns about EVAAS, especially when used for high-stakes consequences in HISD. Dr. Amrein-Beardsley was known in Houston for her role as expert witness in wrongful termination cases of HISD teachers due in part to low EVAAS scores, and she had spoken at HFT hosted events about the strengths and limitations of the EVAAS. As her doctoral student, and by association, district administrators were aware (and weary) of my position and connection to HFT prior to the commencement of this study. That said, there was a lot of room for bias and the burden of proof consistently lain on my shoulders in my role as an objective and fair researcher.

As such, I designed my analytical framework to help position the issues in an academic manner, again following traditional standards used in the field of educational measurement (AERA, APA & NCME, 1999), and to help keep my beliefs in check and in a way that would allow me to objectively examine the realities of EVAAS through the words of teachers in practice. I also carefully constructed survey questions and integrated the feedback I garnered from teachers to better ensure that my own personal biases were not an apparent issue, and that teachers could share their true perspectives and experiences, regardless of their general feelings about EVAAS. As well and as previously stated, I tried to stay as true to the teacher participants' voices and stories to ensure I was not speaking for them, rather them through me. Hopefully each of these approaches help support

my credibility as a researcher and the credibility of the findings presented next in Chapter 5.

CHAPTER 5

Results

In this study I investigated HISD teachers' experiences and perceptions of the EVAAS model used for high-stakes consequences in their district. The main research question was: What are the intended and unintended consequences, as experienced by HISD teachers, through the implementation and use of the EVAAS model? The survey constructs contained questions pertaining to reliability, validity, formative use, and the intended consequences and claimed benefits of EVAAS. As a result of these questions, unintended consequences associated with EVAAS use in HISD were also discovered through analyzing the teachers' responses.

I present the findings following the order listed above, using my analytical framework, again as influenced by AERA, APA, and NCME (1999). First, I describe the teachers in the sample via their demographic and descriptive information. Then, within the reliability, validity, and formative use constructs, I summarize the findings and discuss the main themes as identified through coding the qualitative, open-ended responses. Next, I discuss the findings pertaining to the intended consequences and claimed benefits of EVAAS as measured by the Likert-type scale. Then, I discuss the unintended consequences that I realized as a result of analyzing all of the data throughout the survey. Last, I provide an overall summary of the results that leads into the overall findings and implications warranted and clarified in Chapter 6.

Demographics and Description of Sample

The majority of the teacher participants was female ($n = 648/871$; 74.4%) and identified as Caucasian/White ($n = 306/868$; 35.3%), African American/Black ($n = 237/868$; 27.3%), or Hispanic/Latino(a; $n = 231/868$; 26.6%). The average respondent was 37 years old, with the oldest 78 and the youngest 24.

The plurality of the 882 teachers who responded to the survey had taught in HISD for 6-10 years ($n = 226/878$; 25.7%) and had taught in total for 21+ years ($n = 171/879$; 27.3%). Most of the teachers ($n = 312/882$; 35.4%) had received 5 years of individual EVAAS scores, and overall the average teacher had received 3.64 years of individual EVAAS scores (see Figure 5).

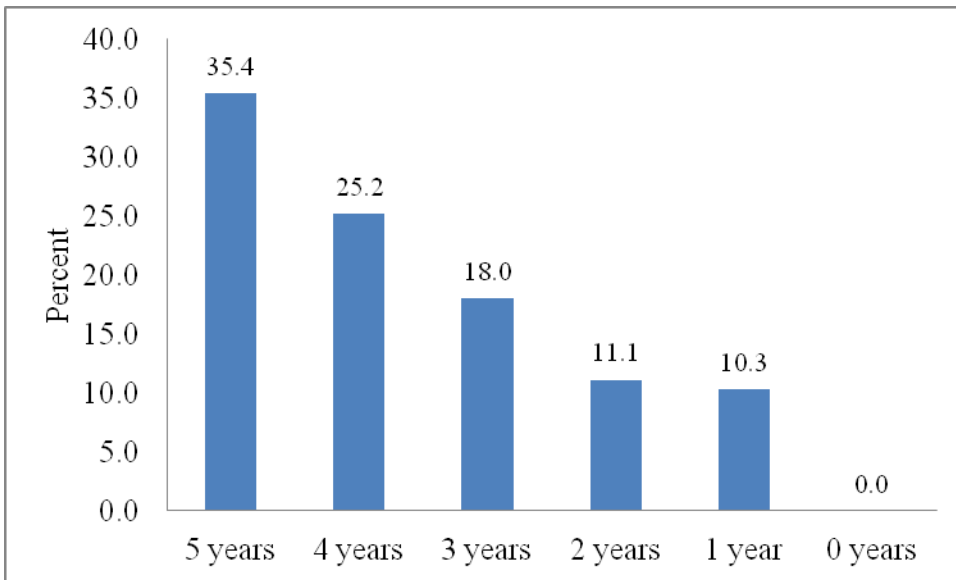


Figure 5. Number of years for which individual EVAAS scores were received.

Most teacher respondents earned their teaching certificates from public universities in Texas ($n = 302/881$; 35.0%) or through alternative certification programs ($n = 266/881$; 30.2%). Almost half of the teachers had taught third ($n =$

412/882; 46.7%) or fourth grade ($n = 417/882$; 47.3%), with the average teacher having taught 3.15 different grade levels (see Figure 6).

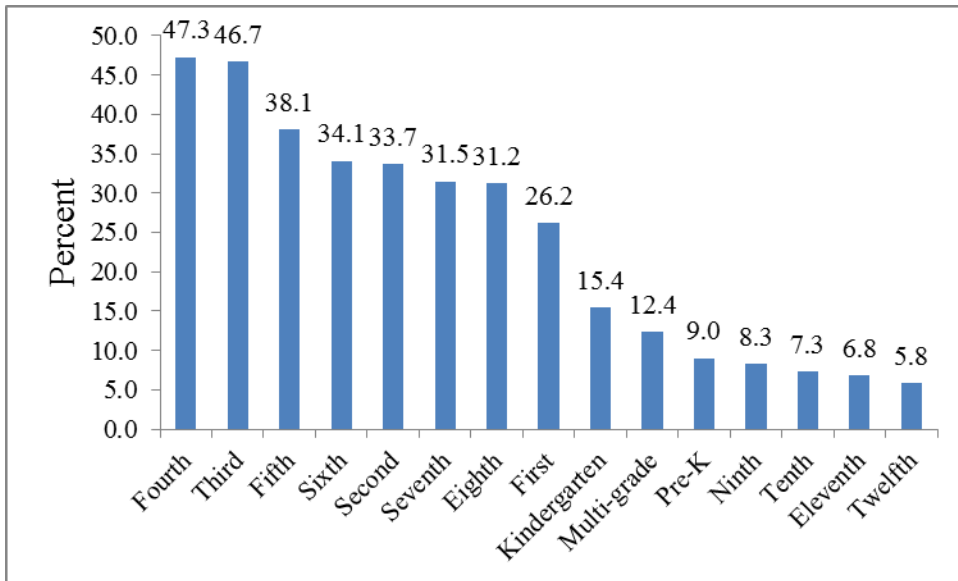


Figure 6. Proportion of grade levels ever taught in the Houston Independent School District.

The majority of teacher respondents taught in the core areas of reading/English language arts, mathematics, science, and social studies, in that order, with the average teacher having taught 3.56 different subject areas (see Figure 7).

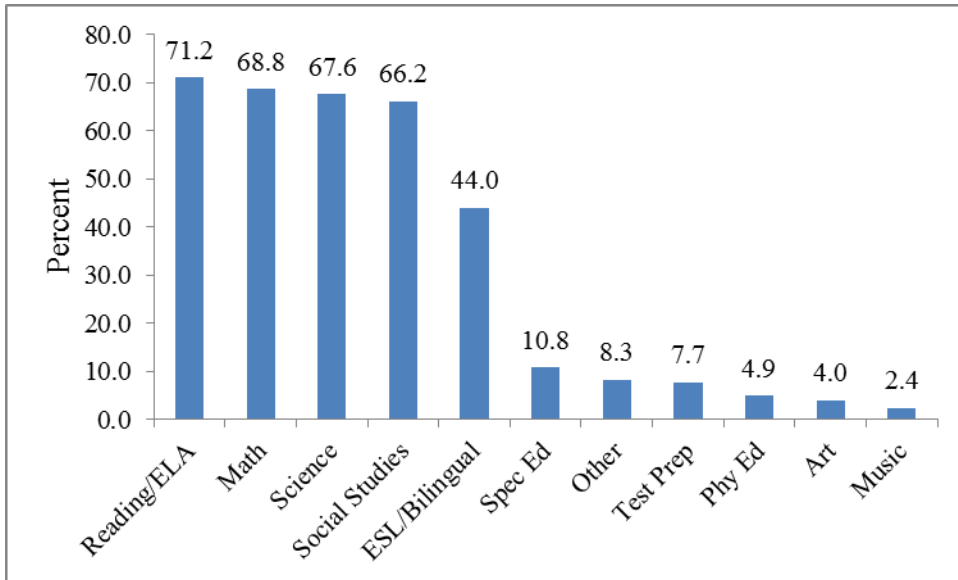


Figure 7. Proportion of subject areas ever taught in the Houston Independent School District.

More than three quarters of the teachers indicated that the students they taught were of high ($n = 692/874$; 79.2%) socioeconomic needs. This makes sense because, again, the majority of the students in the district are from high-needs backgrounds, with 63% of students labeled at risk, 92% from racial minority backgrounds, 80% on the federal free-or-reduced lunch program, and 58% classified as ELLs, Limited English Proficiency (LEP), or bilingual.

Finally, and as mentioned previously, 69.4% ($n = 612/882$) or almost two-thirds of the respondents were members of HFT. Though as the findings later in this chapter show, there was no statistical significant difference among the responses between union and non-union teachers.

Reliability

As discussed in the literature review in Chapter 2 and in the preliminary studies of EVAAS in HISD in Chapter 3, it was evident that inconsistent EVAAS scores year-to-year were an issue of concern. Among participants in this study, more teachers indicated that their EVAAS scores were inconsistent ($n = 404/874$; 46.2%) year-to-year than those who reported consistent scores ($n = 371/874$; 42.4%). Ironically, just like with the reliability construct, their responses were inconsistent, and pretty evenly split. About half of the responding teachers reported consistent data, whereas the other half did not.

To investigate further, I included a question for the teachers who reported inconsistent scores ($n = 404/874$; 46.2%) to provide explanations, and 348 teachers did so generating 381 substantive responses¹⁴ as to why their EVAAS scores were inconsistent year-to-year. Over one third of these respondents ($n = 150/381$; 39.4%) simply replied about how their scores varied, but did not provide explanations for this variation. For example, one teacher replied, “In three years, I was above average, below average and average.” Another teacher responded, “I

¹⁴ Some teachers provided more than one substantive response for each open-ended question, here and throughout the rest of the study. Therefore, the number of substantive responses differs from the number of respondents per question. Superfluous answers (e.g., “No opinion,” “Already answered it in previous question,” etc.) were discarded and not counted as responses or used in qualitative coding.

have taught 4th grade for the last 8 years. My scores have been 'green' some years and 'red' other years.”

Among the teachers who did provide an explanation for the fluctuation of their EVAAS scores, 24.4% ($n = 93/381$) believed the inconsistencies were caused by the different types of students they taught, and (as noted earlier) specifically referenced ELL and transition students as well as high achieving and gifted students as those responsible for the inconsistencies. As one ELL teacher put it, “Since I am teaching 5th grade ELL, I have been categorize[d] as ineffective because my students don't grow when coming from 4th grade all Spanish to 5th grade all English.” A gifted teacher explained:

The first year, they were ok. Then as I began to teach the gifted students, the scores continued to show negative growth. For the 2010-2011 school year, the Principal even told me that my scores revealed that I was one of the worst teachers in the school. The School Improvement Officer observed my teaching and reported that my teaching did not reflect the downward spiral in the scores.

Other teachers ($n = 48/381$; 12.6%) described scenarios of switching grade levels or content areas, which impacted their EVAAS scores as they adjusted to new situations. One new teacher attributed the change in scores to her own growth as a teacher, “My second year's score was higher than my first year's score. I attribute this to professional growth and experience.” Another teacher reflected back on her past four years of EVAAS scores, revealing what she learned along the way:

My first years of teaching I was still learning the ropes. Therefore, those scores were lower; however, over the years I understand that you must teach to the test to get the scores you want. To do well, the students must not only be intimate with the objectives, but also the lay-out and the verbiage on the test. Especially the ELL students. They need to know the wording of the questions beforehand so that they can be sure that they grasp what the question asks.

Reliability across grade levels. To further investigate the issue of reliability, teachers were asked whether their scores were consistent if they had taught more than one grade level. Out of the teachers who reported having taught more than one grade level ($n = 559/873$; 64.0%), 51.3% ($n = 287/559$) reported having inconsistent scores across grade levels. This group of teachers generated 196 follow-up responses to help explain the score inconsistencies across grade levels. The most common explanation came from more than one third of the teachers ($n = 71/196$; 36.2%) who were able to identify a certain grade level that caused their scores to shift. Despite this common response, however, there was no real consensus about the specific grade level responsible for lowering or raising their EVAAS scores. It was also unclear whether decreased scores resulting from moving to a specific grade level was indicative of an adjustment period as the teachers grew more accustomed to the new curriculum, as one teacher noted, “I didn't have enough time to learn the curriculum for the different grade levels I was teaching.”

The next most common response from teachers ($n = 37/196$; 18.9%) was, like just mentioned, that the academic preparation and demographics of students influenced their EVAAS scores. Specifically, they ($n = 23/196$; 11.7%) responded that class size and the proportion of ELL, special education and gifted students in the classroom impacted their EVAAS scores.

A similar number of teachers ($n = 35/196$; 17.9%) indicated skepticism or mistrust about how the EVAAS metrics were calculated, or how the dynamics of a classroom could be quantified or controlled for by a model. As one multi-grade teacher explained:

I did a 7th and 8th grade split one year. 7th grade didn't grow, and [the students] were shown to regress a little bit. 8th grade grew. Was it me? Was it them? Was it both? I tend to think it was them. Chemistry in the classroom can affect the growth, too. And I don't know how that would be measurable by any instrument. Maybe not all things are.

Another teacher whose scores fluctuated in the beginning explained, “When I figured out how to teach to the test, the scores went up.” This also indicates teachers have learned to game the system, or teach to the test, which will be discussed in more detail as an unintended consequence.

Reliability across subject areas. Out of the 577 teachers who had taught more than one subject area, 49.6% ($n = 272/577$) reported that their scores were not consistent across subject areas either. This group of teachers generated 209 substantive explanations for the score fluctuations they reportedly observed. The

two most common reasons provided by teachers were that their scores varied each year regardless of the subject area taught ($n = 74/209$; 35.4%), whereas the same proportion of teachers were able to identify one specific subject area that they believed was responsible for their varied EVAAS scores ($n = 74/209$; 35.4%).

Much like the teachers who believed one grade level was responsible for their EVAAS score fluctuations, however, there was not an overall consensus among this group of 74 teachers as to which subject area(s) caused scores to drop more than others. Although several teachers mentioned that certain subjects such as math and English language arts/reading received more resources. One teacher claimed, “Certain subject areas such as reading and math are given the priority in resources,” and another explained, “My scores tend to be high in math, reading, & writing; but low in science & social studies because we have no or limited materials for those subjects.”

Other teachers described that some subject areas had curricula that were less aligned with the tests than others, specifically those subjects (e.g., history, social studies, and science) that relied on the nationally norm-referenced Stanford test. Additionally, and again, a number of the same teachers ($n = 16/74$; 21.6%) indicated that subjects in which they taught ELL, gifted, and special education students resulted in lower EVAAS growth, regardless of the subject area.

One teacher who could not pinpoint the reason for her score fluctuations recalled, “I receive higher scores in some subjects than in others. Sometimes the most is in my certified field (math) and other times I receive nothing for math but

receive bonuses in other subjects.” Another teacher found better success with her EVAAS scores when she taught history:

When I taught 8th grade history the scores rose significantly one year and stayed consistent for two years. I did nothing different in my approach to teaching. This last year I moved to sixth grade math and the scores took a dip.

Reliability across student characteristics. The final reliability question included in this section of the survey instrument asked teachers if they received consistent EVAAS scores despite the varied proportions of different types of students (i.e., ELL, gifted, special education, low/high income) they taught. Among the teachers who indicated they did teach different types of students year-to-year ($n = 710/877$; 81.0%), 52.5% ($n = 373/710$) responded that their EVAAS scores were inconsistent, yet again. This group of teachers generated 282 substantive responses in explanation.

The plurality of these teachers ($n = 106/282$; 37.6%) responded that all students are different, and that issues such as motivation, prior academic preparation, behavior and external factors such as home life and family situations greatly influenced student performance and inherently teacher EVAAS scores. As one teacher replied, “[EVAAS] depends a lot on home support, background knowledge, current family situation, lack of sleep, whether parents are at home, in jail, etc. [There are t]oo many outside factors – behavior issues, etc.”

Other teachers specifically referenced certain student groups whom they believed were responsible for impacting their EVAAS scores. Gifted and

advanced students were seen by teachers ($n = 49/282$; 17.4%) as high scorers on tests that left little to no room for growth that could be measured by EVAAS. One teacher described working with several different types of students:

GT [gifted], high income students show progress. ELL students usually don't especially if they have taken Aprenda¹⁵ the year before. The Aprenda test really inflates their scores. Fifth graders usually don't show a lot of growth. It seems to be a plateau period for them. EVAAS does not recognize that children's brains might have a plateau period.

And again, teachers listed transition and ELL students ($n = 45/282$; 16.0%) and special education students ($n = 22/282$; 7.8%) as the types of students shown to negatively impact their EVAAS scores as well. This was definitely a recurring theme throughout this section of the study.

Validity

The survey contained several questions to investigate the validity of the EVAAS model and scores as well. To examine content-related evidence of validity, I included questions to determine if the student data used to calculate individual teacher EVAAS scores were appropriate. In addition, I included questions for teachers to compare their EVAAS scores to other indicators of teacher quality to examine criterion-related evidence of validity. The responses

¹⁵ Aprenda is the Spanish equivalent of the Texas Assessment of Knowledge and Skills standardized test used in HISD.

generated from the validity questions overall indicated evidence of an issue with construct-related evidence of validity as well.

Content validity. First, I asked teachers if they had ever been evaluated by EVAAS for a grade level for which they were not the teacher of record. Only 9.1% ($n = 80/875$) of teachers indicated this had happened to them, and these teachers provided 40 substantive responses. The most common explanation for this situation was from teachers ($n = 22/40$; 55.0%) who reported discrepancies with how their teaching responsibilities during student instruction time were allocated to them as part of the data linkage process. Ten teachers ($n = 10/40$; 25%) specifically referenced that the allocation of instruction time split with co-teachers was inaccurate, and four teachers (10%) responded that they had taught more than one grade level in a given year but only received EVAAS scores for students in a certain grade.

A similar minority of teachers ($n = 84/874$; 9.6%) indicated that they had been evaluated with EVAAS scores for a subject for which they were not teacher of record. Sixty teachers generated 57 substantive responses, and the majority of these teachers ($n = 31/57$; 54.4%) indicated they taught in a departmentalized or team-teaching situation or they were a lab teacher, which resulted in inaccurate allocations of student instruction time included in their EVAAS data.

A slight increase was noted when teachers ($n = 152/871$; 17.5%) were asked if they had ever been evaluated with EVAAS scores for students for whom they were not the teacher of record. Of this group, 113 teachers provided 101 substantive explanations for this situation. Almost half of these teachers'

responses ($n = 50/101$; 49.5%) described situations where students were placed in their classrooms only within weeks of the standardized test used to determine EVAAS scores, or where teachers had students removed from their classroom early in the year for disciplinary reasons to attend alternative schools but still had those students' scores show up on their EVAAS reports. A teacher described such a situation:

I'm not sure how I get evaluated for a student who is only in my class for one month and then goes into CEP [community education partners for disciplinary alternative education]. I'm still considered the teacher of record even though he spent 5-6 months out of my classroom.

Other teachers ($n = 31/101$; 30.7%) stated that they were co-teachers, lab teachers or taught in departmentalized classes that resulted in additional students on their roster who were taught by other teachers.

Criterion-related evidence of validity. To examine criterion-related evidence of validity, the teachers were asked if their EVAAS scores typically produced similar findings to their PDAS principal observation scores, assuming both represent accurate measures of teacher quality in HISD. More than half ($n = 497/863$; 57.6%) of the sample indicated their EVAAS scores do not typically match their PDAS scores. Out of this group, 367 teachers generated 340 substantive responses explaining these issues further.

The plurality of teachers ($n = 159/340$; 46.8%) replied that their PDAS scores were always higher than their EVAAS scores, whereas conversely 9.1% (n

= 31/340) of teachers indicated that their EVAAS scores were always higher than their PDAS scores. Regardless of which score was higher, the frequently conflicting EVAAS and PDAS scores seemed to send teachers mixed messages.

One teacher explained:

Based on the EVAAS system, I am considered below the standards, but based on my principal's observation and state test scores, I am a great teacher... Because on one hand you're meeting the State's testing requirements, but if you're [not] doing well according to EVAAS, then you have two contradicting sets of evaluations.

Another teacher responded:

I have always received positive - even glowing - observation and evaluation scores from my principal and evaluator. I have been asked to serve as a lead teacher on campus and I have mentored others - but my negative [EVAAS] growth score does not reflect that.

Other teachers ($n = 43/340$; 12.6%) responded that their PDAS scores were consistent year-to-year while their EVAAS scores fluctuated. Perhaps the consistent PDAS scores result from the more traditional evaluation methods which can lack objectivity. In fact, a fair amount of teachers ($n = 41/340$; 12.1%) indicated that the principal evaluation portion of the PDAS was very subjective, and that principals based their evaluations on their relationships with teachers.

One teacher explained:

If you're 'in' you'll be rated well. If you're not, you won't. The EVAAS scores are nice in that they are purely data driven, and sometimes (if a teacher is [in] a bad way with the principal) they can be a relief.

Some of the same teachers ($n = 41/340$; 12.1%) described how principals would switch their PDAS scores if dissimilar to reflect their EVAAS scores – which are apparently treated as the more important and more objective evaluation score. One teacher said, “Evaluation scores are subjective. One principal told me one year that even though I had high TAKS¹⁶ scores and high Stanford scores, the fact that my EVAAS scores showed no growth, it would look bad to the superintendent.”

Another teacher reflected on when her PDAS scores were changed to match the EVAAS, “I had high appraisals but low EVAAS, so they had to change appraisals to match lower EVAAS scores. I was actually put on a growth plan, but met all the requirements and was taken off.” A veteran teacher explained her changed scores:

One year I received low performing [scores] on my evaluation...I knew the rating was due to her dislike for me. Upon the arrival of the [EVAAS] scores my students did exemplary ...The [assistant] principal changed the [evaluation] rating before I met with her to ‘exceeds expectations.’

¹⁶ Texas Assessment of Knowledge and Skills

To look further at criterion-related evidence of validity, teachers were asked if they had received any awards, recommendations, student or parent feedback, or peer evaluations (again, assuming such indicators also describe teacher quality) which supported or contradicted their EVAAS scores. Out of the teachers ($n = 367/843$; 43.5%) who indicated they had received contradicting feedback, 286 teachers generated 263 substantive explanations. More than a third of these teachers ($n = 95/263$; 36.1%) reported that they had received or were nominated for awards by their colleagues and mentor teachers at the same time they had received low EVAAS scores. Several of these same teachers ($n = 24/263$; 9.1%) pointed out that they were master or lead teachers, department chairs, or development or academic coaches, having been appointed by peers or principals based on their expertise and skill in certain areas, yet they simultaneously demonstrated the “least growth” or had the “weakest” EVAAS scores in the same subject matter.

Other teachers ($n = 81/263$; 30.8%) described the positive feedback they received from parents and students, through letters, personal communication, and continued communication years after students had left their classrooms. Although some could argue these actions are the most subjective of all, for many of the teachers, this feedback served as a more solid indicator of their own effectiveness.

As one teacher put it simply, “Academic testing does not tell the whole story.”

Another teacher shared:

Each year regardless of my EVAAS results, parent[s] request for their children to be in my class. I feel this is because they know I

care about their children and that I am giving them my best each day. Each year my principal must tell parent[s] my class is full.

Formative Use

The literature review in Chapter 2 explained that in order to improve teacher quality, value-added data must be used for formative purposes. Additionally, SAS claims to provide “easily understandable reporting” (SAS, 2012b, p. 1) that can be used by teachers to modify their teaching practices. To understand whether HISD teachers used their EVAAS data in such intended ways, teacher participants were also asked about when, in terms of the academic school year, they received EVAAS reports for their students and in particular for formative use.

The responses indicated there was much variation across HISD, likely from school-to-school, in the distribution of EVAAS reports to the teachers. Understanding that the exact timing of standardized test administration varies slightly year-to-year, teachers were asked to select all responses that applied to their experience receiving EVAAS reports. The majority of teachers ($n = 530/882$; 60.1%) indicated that they received their EVAAS reports in the summer or fall, when the students responsible for generating EVAAS information were no longer under their instruction or they were in the next grade level. Just over 10% ($n = 107/882$; 12.1%) of teachers responded that they received EVAAS reports for students prior to the students entering into their classrooms. Additionally, some teachers indicated that they typically do not receive individual EVAAS scores for

their students ($n = 88/882$; 10%), whereas others reported never having received individual EVAAS scores for their students ($n = 51/882$; 5.8%) (see Figure 8).

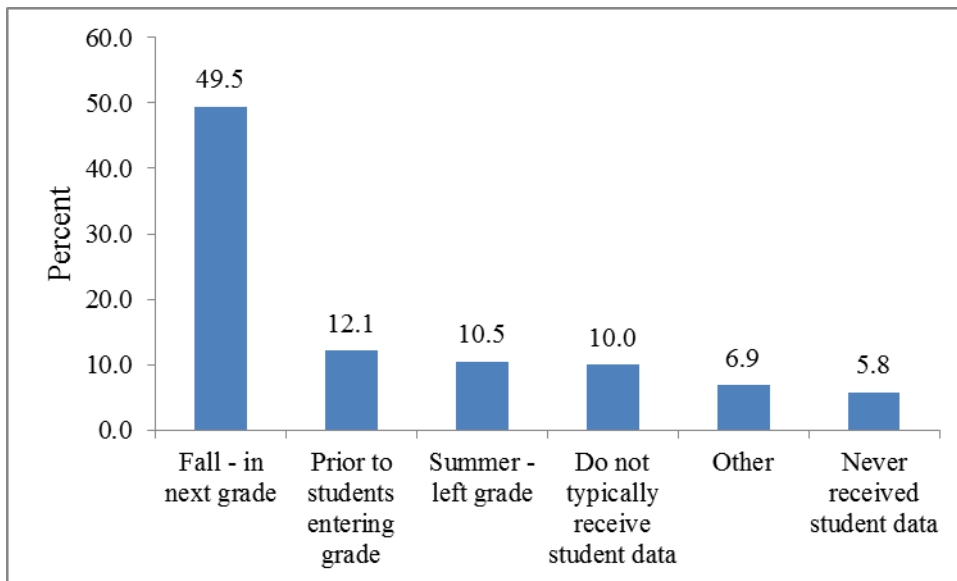


Figure 8. When teachers typically receive EVAAS reports for their students.

Next, teachers were asked if they had received EVAAS reports for their students, whether they used the information to inform their instruction. The majority ($n = 480/815$; 58.9%) indicated they do not use EVAAS reports to inform their instruction. The teachers ($n = 335/815$; 41.1%) who indicated that they did use EVAAS data, were asked to explain how. Out of that group, 222 teachers provided 238 substantive explanations for how they used EVAAS data to inform their instruction.

The most common response was from teachers ($n = 53/238$; 22.3%) who responded that they knew they were “supposed to” look at their EVAAS reports, so they would look at the reports to get an overview on how the students performed; however, teachers called the reports “vague” and “unclear” and were

“not quite sure how to interpret” and use the data to inform their instruction. As one teacher explained, she looked at her EVAAS report “only to guess as to what to do for the next group in my class.” Another teacher responded:

[I] attempted [to use the them] but the reports are not helpful at all. They are a mixture of Stanford and TAKS. I need to know what the anticipated TAKS and Stanford scores are so I can make goals for myself and [my] students; however, since part of EVAAS [is] comparing teachers at like schools, the goal is a moving target.

A third teacher added that the timing of report distribution prevented her from using the EVAAS data, “By the time I get the scores the students are in another grade. I can look at the previous years’ scores, but [the reports] have to be pulled by individual students...This is too time consuming.”

Other teacher participants ($n = 45/238$; 18.9%) described how they used their EVAAS reports, stating that they used the reports for ability grouping to differentiate instruction, whereas, related, others ($n = 44/238$; 18.5%) reported that they looked at the EVAAS reports to plan for remedial instruction with future students. One teacher explained, “If I’m low in one area, I try to maximize instruction in that area...I was low in [social studies] so I tried to incorporate more [social studies] activities into reading lessons.”

However, even among the teachers who indicated that they used EVAAS for ability grouping, differentiating instruction, and remedial education, very few actually articulated *how* the data were specifically used. This teacher started

describing how she used EVAAS reports to look at subgroups, but then revealed she was not quite sure what to do beyond that:

EVAAS is most helpful for me when looking at subgroups and their growth. For example, you can look at the growth of just boys, or girls in your class. You can also look at gifted versus non-gifted. I believe looking at how each subgroup performed is helpful. The only issue is that you're not 100% sure how this score is calculated, so it's not clear what part of your practice you should go back and change.

Another teacher responded, “I do use them, but only to tell me what level these students are on and how much growth they need to make. It is not specific enough to tell me exactly the strength or weakness in each area.”

Other teachers ($n = 24/238$; 10.1%) indicated that they use EVAAS reports to identify the lowest performing kids to pull out for tutoring or remediation, and also the “bubble kids” whom they usually focused their teaching efforts on to try to maximize growth scores. Teachers identified bubble kids as students who performed just below average, with greater relative potential to demonstrate EVAAS growth. As one teacher who used his EVAAS reports in this way explained, “It allowed me to focus on my bubble students early.” A handful of other teachers ($n = 15/238$; 6.3%) who indicated that they used EVAAS reports to inform their instruction responded that they actually used other data to inform their instruction instead, such as those derived via the Stanford and TAKS.

I included an additional question about formative use on the survey, but this time provided teachers with the opportunity to make multiple selections to describe which of the listed scenarios described their EVAAS data usage best (see Table 5).

Table 5

EVAAS Data Usage by Teachers

Multiple Selection Scenarios for EVAAS Data Use	<i>N</i>	Percentage
You use other resources (not EVAAS data) to inform practices	400/882	45.4%
You use EVAAS data to inform your classroom practices	242/882	27.4%
You do not typically use EVAAS data to inform your practices	220/882	24.9%

As illustrated, the most common response among teachers (45.4%) was that teacher respondents' used other resources (not EVAAS data) to inform their teaching practices. Just over a quarter of the teachers responded that they use EVAAS data to inform their practices, and almost a quarter of the teachers responded that they did not typically use EVAAS data to inform their instructional practices whatsoever. Here, it seems, teacher respondents valued other data more than EVAAS, and responses were mixed regarding whether respondents valued EVAAS data in general.

Last, teachers were asked to describe the extent to which they typically reflected on their EVAAS reports to improve instruction, and ($n = 559/882$; 63.4%) teachers generated 476 substantive responses. The most common response among teachers ($n = 156/476$; 32.8%) was that they do not use EVAAS reports to

improve their instruction, and many ($n = 53/476$; 11.1%) of these same teachers further explained that they “do not understand” or “don’t trust the EVAAS data.” Another group of teachers ($n = 61/476$; 12.8%) replied that they tried to use their EVAAS reports but either could not understand them, or by the time they received the reports the students had left their classroom and the variation of students in each class prevented the application of information from one group of students onto the next.

Among the teachers ($n = 86/476$; 18.1%) that indicated they did reflect on EVAAS data to inform their instruction, they, again, described situations of ability grouping and using student projections to plan lessons to meet certain student needs. Some teachers were able to better articulate how they used their data than others. For example, one teacher said, “I would look for patterns in the data regarding what types of students made the most/least growth. I would also compare EVAAS growth to my own records of their achievement (number of books read, grades, etc.)” Another said, “Their individual data drives my instruction. I use it to build individual study guides based on their needs. I use their testing history and personal observations.” Interestingly, the teachers who explained how they used EVAAS data also reported using their EVAAS data along with other measures of student growth and performance – not relying solely on EVAAS for formative use.

Finally, another group of teachers ($n = 42/476$; 8.8%) replied that they did use EVAAS scores to inform practices, but again provided vague descriptions, providing responses like “yes” and “I use it.” An additional 38 teachers (8.0%)

responded that they used their EVAAS data “all the time” or “constantly” but also did not provide further explanations for how they used the data.

Formative Use Support

Shortly after HISD contracted with SAS, the district began offering both in-person and online training courses for teachers to learn about the EVAAS model and its resultant data (HISD, 2010). When asked about these trainings and opportunities, just over one-third of the teachers ($n = 324/870$; 37.2%) indicated that they were unaware of EVAAS training sessions provided by the district to help teachers understand the model and reports. The plurality ($n = 404/863$; 46.8%) of teachers reported such trainings were optional and 23.1% ($n = 199/863$) indicated the trainings were mandatory.

Of those teachers who were reportedly aware of such in-person ($n = 546/882$; 61.9%) and online ($n = 338/882$; 38.3%) training sessions, the teacher respondents indicated the number of sessions they attended (see Table 6).

Table 6

EVAAS Training Session Attendance

Number of Sessions Attended	In-Person	Online
1	(<i>n</i> = 195/546) 35.7%	(<i>n</i> = 160/338) 47.3%
2	(<i>n</i> = 181/546) 33.2%	(<i>n</i> = 98/338) 29.0%
3	(<i>n</i> = 91/546) 16.7%	(<i>n</i> = 47/338) 13.9%
4	(<i>n</i> = 31/546) 5.7%	(<i>n</i> = 28/338) 8.3%
5+	(<i>n</i> = 48/546) 8.8%	(<i>n</i> = 55/338) 16.3%
Total	(<i>n</i> = 546) 100%	(<i>n</i> = 338) 100%

As illustrated, more teachers reported having attended in-person EVAAS training sessions than online; however, 62.1% (*n* = 386/622) of the teachers found such trainings unhelpful in terms of helping them better understand the EVAAS model and scores. Perhaps that is the reason the majority of sampled teachers did not attend another training session.

To investigate further, teachers were also asked if their principal or supervisor typically discussed their EVAAS results with them. Slightly more teachers (*n* = 422/868; 48.6%) responded that their principals did discuss their EVAAS results with them than those teachers (*n* = 397/868; 45.7%) who did not discuss their EVAAS results with a principal or supervisor. However, in analyzing the 277 substantive explanations provided by those teachers who had

discussed EVAAS with their principals, it became clear that not all teachers had similar experiences or discussions.

The most common explanation of such circumstances came from the teacher respondents ($n = 85/277$; 30.7%) who indicated that their principals told or showed them their scores in a manner that was “vague,” “not in depth,” and “not discussed thoroughly.” Of these 85 teachers, 31 specifically indicated that they thought the “very basic discussions” were due to the fact that their principals did not understand the EVAAS reports either. One teacher explained, “He looks at them [EVAAS scores], but is unable to explain them.” Another teacher stated that his principal “goes over the data, without much comprehension on how scores are derived. [The principal] cannot suggest improvements.” Another teacher replied, “Our principal does not know how they get the score and has tried many times to get someone to come and explain it to us. No one can.”

The next most common description of such circumstances was provided by teacher respondents ($n = 56/277$; 20.2%) who reported that their principals discussed their EVAAS reports with them at the end of the year during performance evaluations, but teachers did not provide much explanation for these discussions. One teacher said their principal discussed EVAAS “during the last conference together at the end of the year. [I] would like better feedback/support in how to improve.” Another teacher said the EVAAS reports were discussed, “At the end of the year, before we can get hired again.”

A similar number of teachers ($n = 51/277$; 18.4%) indicated that their principals discussed their EVAAS scores in a group setting or team discussion,

but not individually with each teacher. Other teachers ($n = 33/277$; 11.9%) reported that their principals discussed their EVAAS reports with them at the beginning of the year to set yearly goals. But out of all the responses, only 4.7% ($n = 13/277$) of the teachers reported that their principals were able to “explain what the scores mean” or tell teachers “how to use the data to improve scores.”

Intended Consequences and Claimed Benefits of EVAAS

The final section of the survey was designed with items meant to gather teacher participants’ perspectives on the intended uses (consequences) and claimed benefits of EVAAS, as well overall EVAAS statements generated to further capture teacher perception of the model. The same Likert-type scale was used to capture teachers’ levels of agreement with the following statements with values, again assigned as: Strongly Agree (SA) = 5, Agree (A) = 4, Neither Agree nor Disagree = 3, Disagree (D) = 2, Strongly Disagree (SD) = 1 (Gay, 1996, p. 155; see Table 7).

Table 7

Items Capturing Respondents' Opinions about EVAAS Statements

	Statement	<i>N</i>	<i>M</i>	<i>SD</i>
2	EVAAS helps create professional goals	870	2.27	1.25
3	EVAAS helps improve instruction	864	2.24	1.23
11	EVAAS will provide incentives for good practices	860	2.19	1.24
5	EVAAS ensures growth opportunities for very low achieving students	875	2.15	1.18
4	EVAAS ensures growth opportunities for students	873	2.14	1.16
7	EVAAS helps increase student learning	868	2.13	1.16
8	EVAAS helps you become a more effective teacher	869	2.12	1.21
15	Overall, the EVAAS is beneficial to my school	855	2.10	1.22
1	EVAAS reports are simple to use	866	2.09	1.14
14	Overall, the EVAAS is beneficial to me as a teacher	858	2.08	1.25
16	Overall, the EVAAS is beneficial to the district	847	2.08	1.23
6	EVAAS ensures growth opportunities for very high achieving students	870	2.06	1.14
10	EVAAS will identify excellence in teaching or leadership	849	2.00	1.15
9	EVAAS will validly identify and help to remove ineffective teachers	849	1.88	1.10
12	EVAAS will enhance the school environment	842	1.86	1.11
13	EVAAS will enhance working conditions	842	1.76	1.04

Note. Items are arranged by *M* in descending value.

The descriptive statistics above illustrate that all mean values were between 1.76 and 2.27, which indicates that the average teacher disagreed more than they agreed with each of the EVAAS statements presented to them in this section of the survey instrument. In fact, more than 50% of the teachers disagreed

or strongly disagreed with every single statement, and less than 20% of the teachers agreed or strongly agreed with every statement.

The teachers disagreed most with statement 13, “EVAAS will enhance working conditions” with 75.7% ($n = 637/842$) of the teachers disagreeing or strongly disagreeing with this assertion. Similarly, 72.9% ($n = 619/849$) of the teachers disagreed or strongly disagreed that “EVAAS will validly identify and help remove ineffective teachers,” and only 72.5% ($n = 611/842$) of the teachers disagreed or strongly disagreed or agreed that “EVAAS will enhance the school environment.” A table for all Likert-type items responses, including the number and proportion of respondents for each statement can be found in Appendix E.

As mentioned throughout this study, there was a high volume of HFT members ($n = 612/882$; 69.4%) represented in the sample. As such, I calculated chi-square analyses for each of the survey questions with categorical responses to examine whether the perceptions among the two mutually exclusive groups, union and non-union teachers, differed at statistically significant levels (Gay, 1996). Table 8 represents results from the chi-square analysis for Statement 1 below.

Table 8

Chi-square Analysis for Statement 1

Statement and Chi-square result	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Total
1: EVAAS reports are simple to use; $\chi^2 = (4, N = 866) = 1.96, p = .744$						
Non-HFT	111 (12.8)	74 (8.5)	39 (4.5)	39 (4.5)	5 (0.6)	$n = 268$ (30.9%)
HFT	250 (28.9)	147 (17.0)	107 (12.4)	83 (9.6)	11 (1.3)	$n = 598$ (69.1%)
Total	361 (41.7)	221 (25.5)	146 (16.9)	122 (14.1)	16 (1.8)	$n = 866$ (100%)

Table 8 shows that the chi-square value of 1.96 with 4 degrees of freedom is not significant at conventional significance levels ($p = 0.744 > 0.05$). This signifies that there is no statistical difference between non-union and union teachers on their agreement (or disagreement) with the statement that EVAAS reports are simple to use. In fact, none of the numerical statements included in the survey instrument yielded statistically significant differences between union and non-union members' responses ($p < 0.05$). For all related results, see Appendix F for chi-square tables for statements 2-16 above, and *see* Appendix G for all other chi-square tables pertaining to all other numerical items included in the survey instrument.

The last structured question on the survey instrument in this section asked teachers what and who they would select to teach and why, if their sole purpose was to gain the highest EVAAS scores. Teachers who responded to this question ($n = 597/882$; 67.7%) generated 601 substantive responses, capturing a wide

range of different aims. The most common response came from 11.8% ($n = 71/601$) of the teachers who responded that they would select “low achieving” or “academically needy kids” who performed poorly on previous grade level assessments who had “nowhere to go but up.” One teacher added, “We call these students our money makers.”

A similar number of teachers ($n = 68/601$; 11.3%) responded that they would select to teach math. Several of these teachers indicated that math was not necessarily dependent on the English language and therefore, it was perceived that the EVAAS scores would not be impacted by transition, ELL, or bilingual students. Related, teachers ($n = 46/601$; 7.7%) listed specific groups of students they would select not to teach to receive the highest EVAAS scores, which included ELL, transition, bilingual, special education, and gifted students, as teachers, again here and elsewhere throughout the study, indicated that these students typically showed little or no growth as measured by EVAAS and, accordingly, weigh or bias teachers’ composite EVAAS scores downwards.

Other teachers ($n = 54/601$; 9.0%) stressed things like wanting to teach students from middle to high socioeconomic status backgrounds, who they perceived as more likely to have parents who believed in discipline and were relatively more engaged in their children’s learning processes. Another group of teachers ($n = 40/601$; 6.7%) responded that they would teach “bubble kids” or average or middle-range students who demonstrated that they were capable of learning and still had room to grow. Almost the same number of teachers ($n = 37/601$; 6.2%) responded that they would teach an early grade level (K-3) or a

non-tested subject, largely to avoid the accountability pressure prevalent in the district.

Other teacher respondents specifically mentioned other subjects they would teach: English/language arts or reading or writing ($n = 37/601$; 6.2%), social studies or history ($n = 33/601$; 5.5%), and science ($n = 30/601$; 5.0%). It was unclear, however, whether the teachers who stated that they would select to teach a specific subject currently taught the subject area selected, but some added notes such as “it is my best subject” or “that is what I am certified in” and “I enjoy teaching it.”

Otherwise, 3.5% ($n = 21/601$) teachers indicated they would leave HISD or quit teaching all together because they “don’t teach for the EVAAS scores,” they teach because they are passionate about the subjects and students they teach. As one teacher stated, “I never needed an incentive to teach to do my best, I love to teach.”

Unintended Consequences

Through analyzing teachers’ responses to the survey questions previously discussed, many unintended consequences were also discovered via the aforementioned analyses that seem to be occurring as a result of EVAAS implementation in HISD as well. Such unintended consequences include: disincentives for teaching certain student groups (as mentioned above); teacher mobility issues with teachers looping or teaching back-to-back grade levels and switching grade levels within the same content areas (as mentioned in the focus group discussions highlighted in Chapter 3); cheating or teaching to the test as a

result of accountability pressures (as evidenced in the general literature about high-stakes testing and also Chapter 2); general distrust of the EVAAS model (as also mentioned, particularly in terms of transparency), competition and low morale among teachers, both of which are foreseen as perceived effects from EVAAS implementation (as also evident in the general research literature about high-stakes accountability systems).

Disincentives for teaching certain groups of students. Again, as evidenced throughout the study, teachers identified working with certain populations of students as problematic if they are to achieve high EVAAS scores. Specifically, high performing and gifted students who are inhibited by ceiling effects, transition students who are in their first year of English-only instruction, and teachers in classrooms with high proportions of special education and ELL students were of exceptional concern. As one teacher said, “it is extremely difficult to raise test scores for [gifted] students.” Another teacher described her frustration with low EVAAS scores, stating that she is being “punished for teaching ELL and [gifted] students.” If it were possible, the same teacher noted, “I would refuse to teach ELL and [gifted] students.”

Another teacher, certified to teach ELL students, described her experience with EVAAS scores, as a teacher of both ELL and gifted students:

The more ELL students I have, the lower my scores. I also have [gifted] students that [sic] score high in 2nd grade on the norm referenced test and their scores are compared with the 3rd grade criterion referenced test so there is no room to show growth.

Another teacher discussed how she had the most difficulty obtaining high EVAAS scores when teaching special education students:

I had 11 special [education] kids last year with no co-teacher [or] assistance of any kind. The kids' disabilities ranged from emotional disturbances to learning disabilities to borderline retardation. I had a higher failure rate with them than with my other classes.

Teacher mobility issues. Though EVAAS claims to be able to account for teacher mobility, results from this study also provide evidence indicating that in HISD it is pretty common to loop, or teach the same content area in back-to-back grade levels. When these teachers have the same students within their classrooms for back-to-back years, teacher respondents expressed difficulty showing EVAAS growth two years in a row.

One teacher noted, “My scores have always varied from the absolute highest to the absolute lowest, even when I taught the same exact kids two years in a row.” Another said, “I teach many of [the] same students in 7th and 8th [grades]. In 7th I show growth, then in 8th [I] suffer.” A different teacher described how she used drill and kill test preparation to reverse the looping effect on her EVAAS scores:

In 2nd grade my students scored so high (90th percentile), it was almost impossible to show growth with the same students in third [grade]. After realizing this, the next year in third [grade] I gave my student[s] twice as much test prep as I had the year before when they did not show any growth, preparing them for tricky

multiple choice questions. The result was outstanding! I received a huge bonus and showed so much growth, but sad to say [sic] because of more test prep.

Other teachers described the flip-flop effect (as described in Chapter 3; see also Amrein-Beardsley & Collins, 2012), whereas a teacher rated as effective by EVAAS would switch grade levels with an ineffective teacher, and his/her EVAAS ratings would flip-flop from the mere move. One teacher explained, “When I taught 8th grade they were very different from when I moved to 6th grade.” Another teacher reported, “I taught social studies to my [5th] grade homeroom class and I was below district expectations. Previously, when I taught it to 4th graders I was considered above expectations.”

Cheating and teaching to the test. As a result of the pressures teachers are under to obtain high EVAAS scores, some respondents also indicated that there was evidence of cheating and other unprofessional and unethical behaviors occurring as principals and teachers tried to increase their EVAAS scores. The various scenarios described by teachers spanned from befriending principals to hand-select their own class rosters to admitting to a drill and kill teaching approach to ensure students know the material for the high-stakes tests. One teacher claimed that, “EVAAS is creating a very competitive setting. The teachers want to recruit the best profiles. There are conversations ‘during the summer’ to obtain the best rosters.” Another teacher described the opposite scenario with principals, “If they don't like you they stack [your roster with] the students with issues, give you no support and crucify you with EVAAS. It's a set up.”

As a result, some teachers expressed the activities and behaviors of principals as “corrupt.” As one teacher stated, “I think our district is just trying to get means to get rid of teachers due to a system that nobody understands, but for HISD, [EVAAS] is a legal way to do it.” Another teacher said, “It doesn't matter if you're a good or bad teacher; if your principal or supervisors want you out then you will be out. They can put anything on your evaluations.” A different teacher responded, “The principal is the one who has the last word concerning EVAAS.”

As referenced by teachers throughout the findings of this study, teachers discussed “drill and kill” teaching approaches, “teaching to the test,” and reported knowing that “teachers cheat” to increase their EVAAS scores. One teacher explained, “If [two] or more teachers can work together to cheat with each others' students then they stand to profit \$7,000-\$10,000 per school year. That's upwards 3 times what could be made teaching one session of summer school.” Another teacher claimed, “You must be willing to teach strictly to the test, or be willing to cheat because that is the only way your [EVAAS] data will look good.” Yet another described, “To gain the highest EVAAS score, drill and kill and memorization yields the best results, as does teaching to the test.”

Numerous teachers reflected on their own questionable practices. As one teacher said, “When I figured out how to teach to the test, the scores went up.” Another added, “Anything based on a test can be ‘tricked.’ EVAAS leaves room for me to teach to the test and appear successful.”

However, teachers were also cognizant of the negative impacts that teaching to the test has on students. A veteran teacher claimed, “There is no real

teaching anymore because the scores obsession is driving teachers to teach to the test. Students are learning to bubble an answer sheet instead of learning to think and reason.” A math teacher expanded on this idea:

As a result of the emphasis on EVAAS, we teach less math, not more. Too much drill and kill and too little understanding [for the] love of math. Students who have come up with this in place are remarkably less likely to think and inquire and innovate, and more likely to sit-and-get. Raising a generation of children under these circumstances seems best suited for a country of followers, not inventors, not world leaders.

Another teacher took these concerns one step further and discussed students’ long-term well-being and success:

When they arrive at college, they are unprepared to write, read, take higher level assessments since the curriculum has been "dumbed down" to make sure that the students pass, and whatever the cost. Students today have been enabled so much that when they get to college, they are shocked when they flunk out because they can't retake a test until they finally pass...I strongly feel that today's students are not expected to perform at the standards they need to be performing at...In other words, I would not want to get on an airplane that was piloted by one of my students, would not trust any of them to be my physician or attorney since I would be

afraid that they were just "passed on" to the next grade level by the public schools...

Distrust, competition, and low morale. Lack of transparency surrounding the EVAAS model and data appears to have created a sense of distrust among teachers as well. A teacher shared her sentiments:

Ultimately, there are no stated metrics and as such I don't trust that the people who assign this number are using this in my or my school's best interest. To use the lingo, the current system is not transparent. That makes me more resistant to data [or] a system that has the potential to be very useful for testing.

One teacher acknowledged the sophistication of the EVAAS metrics, but added that he was skeptical of its usefulness, "I don't completely believe in it or trust that the calculations are valid. And even if the whole EVAAS operation is mathematically sound, I'm still not sure if it is all that important." Another added, "Since I don't find the reports consistent with my instruction, effort and quality of practice, I don't trust EVAAS reports."

Another teacher expanded on how distrust has impacted teacher collaboration and may be harming students:

Since the inception of the EVAAS system, teachers have become even more distrustful of each other because they are afraid that someone might steal a good teaching method or materials from them and in turn earn more bonus money. This is not conducive to having a good work environment, and it actually is detrimental to

students because teachers are not willing to share ideas or materials that might help increase student learning and achievement.

Otherwise, it seems that the ASPIRE bonuses attached to EVAAS output have also lowered morale and created a sense of competition among teachers. Numerous teachers reported this as an unintended consequence, and one teacher captured these teacher respondents' views best by noting, "It [EVAAS] trades 'it takes a village' for 'every man for himself.'" Another mentioned, "This system undermines collaboration, a cooperative work place, and pits administration against the staff." Yet another teacher referenced the competition that has emerged as a result of using EVAAS scores for the ASPIRE program by writing:

The ASPIRE incentive program is not an incentive. For something to be an incentive, you need to know what you have to do to get the incentive. All we know is that as a teacher you have to improve your scores more than the other teachers. You can make improvements each year, but if other teachers improve the same amount, you have made no gains according to the system. It is a constantly moving target. You don't know what you need to do to get the "prize" until after the "contest" is over.

Another teacher described her own weakened morale and how even non EVAAS eligible teachers in HISD have been impacted. She wrote:

EVAAS and the bonuses attached to it are tearing down the morale of our school. Before, we worked as a team to get our kids where they needed to be. The first year that EVAAS and ASPIRE came

around, I got a school-wide \$500 ASPIRE payment while the science teacher down the street got around \$6,000. The next time I was asked to take a leadership role with science, I told them to ask the lab teacher down the street...obviously they knew more about getting results than I did. After a couple of weeks, I calmed down and took on the responsibility offered me, but I still resented the unfairness of it all. I see the same attitude with our lower grade teachers. They feel like they are chopped liver compared to the testing grades. We need tutors to help out with our struggling kids in the testing grades, and usually we rely on our lower grade teachers to help out. This year, we can't beg, borrow, or steal anyone to stay after school or come in on Saturdays. Our upper grade teachers are barely running on steam, and our lower grade teachers feel unappreciated and disenfranchised, and say, "They're getting the big bucks, let them earn it." This is not a business, this is education. There is no formula or secret recipe that is fool-proof. These are kids. They are unpredictable, extravagant, ignorant, brilliant, talented, clumsy, graceful, insightful, and clueless... all wrapped up in one. They don't even know themselves yet.... and my career depends on how they do on a test that they take one day out of the 3 years that I teach them science?

Summary of Results

This chapter presented the results from the main research question, highlighting the intended and unintended consequences of the EVAAS as experienced and perceived by the teachers in HISD. In this chapter, I also presented the issues as expressed by teachers, with regards to issues with reliability, validity, and the formative uses of EVAAS data, as well as teacher perceptions about various EVAAS statements and marketing claims. A series of unintended results that also emerged throughout data analyses were also illustrated. Next, the overall findings as inferred from these results, along with their implications for multiple policy and practitioner audiences, overall conclusions, and recommendations for further research are discussed in Chapter 6.

CHAPTER 6

Findings and Conclusions

In this chapter, I conclude the dissertation by summarizing the study and discussing the overall findings and implications inferred from the results, as additionally supported by, and resituated within the literature. Last, I discuss overall conclusions and provide recommendations for further research.

Summary of the Study

Missing from the research field are studies that examine value-added beyond just the value-added scores and the statistics used to derive them. Void are studies in which researchers look at whether value-added output are representative of the characteristics (i.e., “ineffective” or “effective”) associated with teachers receiving such scores (Hill et al., 2011). Also missing are studies that examine how teachers and their practices have been impacted by value-added methodologies and accountability policies based at least in large-part on value-added output. In this study I sought to understand whether teachers use value-added data to reflect and improve upon their instruction, mainly to bring the invaluable perspective and voices of the teachers into the national conversation. This is the first study that has taken this approach.

HISD has contracted with SAS to use the EVAAS model to evaluate teacher effectiveness since the 2006-2007 academic year. EVAAS data are used by HISD to determine merit pay as part of their ASPIRE program, and EVAAS output can ultimately impact termination decisions. Because HISD uses value-added data more than any other district or state in the country for high-stakes

purposes (Corcoran, 2010; Harris, 2011; Mellon, 2010; Otterman, 2010; Papay, 2010), the purpose of this study was also to gain the aforementioned understandings (i.e., regarding how teachers and their teaching practices have been impacted by the implementation and use of the EVAAS model to hold them accountable) in the district that uses these data the most and in the most consequential ways. In this study I also examined teachers' perceived realities about the model's advertised utility, as marketed and advertised by SAS.

My overarching research question was: What are the intended and unintended consequences, as experienced by HISD teachers, through the implementation and use of the EVAAS model? The survey research study contained four constructs as aligned with my analytical framework, with sub-questions regarding: (a) Reliability – Are EVAAS scores consistent over time? (b) Validity – Do EVAAS scores match other indicators of teacher quality? (c) Formative use and consequences – Do teachers use EVAAS data to inform their instruction? (d) Intended consequences and claimed benefits of EVAAS – Do teachers agree with EVAAS marketing claims and statements? I also included demographic questions pertaining to participants' years of teaching experience, subject area(s) taught, grade level(s) taught, where they received teaching certification, and (non)union affiliation.

In my attempt to include the perspectives and experiences of as many EVAAS eligible HISD teachers as possible, I used a large-scaled online survey instrument to facilitate my survey research approach. Only those teachers who indicated that they had taught in EVAAS eligible grade levels 3-8, and had

received one or more year of individual EVAAS scores, were included in the final analyses. The actual response rate probably fell somewhere between the range of 14.0%-20.1%. The exact response rate could not be determined, again, as HFT also distributed the survey, and the email list I used included a fair amount of HISD teachers who were EVAAS ineligible. Therefore the total number of teachers who received the survey and the denominator were unknown.

I analyzed all data using a mixed methods approach, using the qualitative data to support the quantitative data and vice versa. This approach allowed for more insight and greater comprehensive findings that I believe are easy to understand and user-friendly, and of course that could also be compared to and resituated within the literature on the topic.

In the previous chapter, the results were presented as aligned with my analytical framework: reliability, validity, formative use, and intended consequences and claimed benefits of EVAAS. Additionally, unintended consequences associated with the implementation and use of EVAAS emerged throughout the teacher participants' responses to the survey questions.

In the next section, I present the overall findings as based on these results for each construct. Within each construct, I will discuss implications for both HISD and the larger educational audience, including policymakers and practitioners. Last, I present recommendations for further research in this area.

Overall Findings and Implications

Reliability. According to teachers who participated in this study, reliability as measured by consistent EVAAS scores year-to-year was ironically,

an inconsistent reality. About half of the responding teachers reported consistent data whereas the other half did not, just like one would expect with the flip of a coin (see also Amrein-Beardsley & Collins, 2012). Similarly, teachers reported split consistencies of EVAAS scores across grade levels and different subject areas taught (LeClaire, 2011), as well as given varied student characteristics (Hill et al., 2011; Newton et al., 2010; Rothstein, 2009). Continuing with the inconsistent data, about half of the teachers claimed their EVAAS scores fluctuated regardless of the grade level or subject they taught, whereas the other half indicated one specific grade level or subject was responsible for their lower EVAAS scores. However, there were no apparent trends related to the particular grade level or subject area caused the scores to drop more than others. These fluctuations resulted in teacher effectiveness classifications shifting year-to-year, although teachers believed their teaching techniques remained consistent. Such reliability issues and misclassifications are becoming more noted in literature as well (Baker et al., 2010; Haertel, 2011; Koedel & Betts; 2007; Papay, 2010).

Related, teachers who reported EVAAS score inconsistencies identified students as the main cause for the score fluctuations they observed, regardless of grade level or subject areas taught. These teachers specifically mentioned the impact that motivation, behavior, prior academic preparation and demographic influences such as family support and home life all have on EVAAS scores, which is contrary to what EVAAS creators indicate can be “statistically controlled for” (see Sanders & Horn, 1998).

Throughout all of the reliability questions, the consensus among teachers was that gifted, transition, ELL, and special education students were the most difficult student groups to demonstrate high levels of growth as measured by EVAAS. Even with the most sophisticated controls and blocks, it appears that EVAAS cannot control for the impact of extraneous variables such as home life, health, behavior, motivation, etc. on student learning (see also Haertel, 2011; Harris, 2011; Rothstein, 2009). Additionally, more than half of the HISD teacher participants indicated they had taught more than one grade level, more than one subject area, or different types of students each year, yielding inconsistent EVAAS data at least 50% of the time.

Reliability implications. Unless a school district could prevent teacher mobility and ensure equal, random student assignment, it appears that EVAAS is unable to produce reliable results, at least greater than 50% of the time. As such, it is highly inappropriate and invalid for HISD (and any other district) to use unreliable EVAAS results for anything since a teacher seemingly has the same probability of being rated “effective” or “ineffective” as (s)he would calling “heads” during a coin toss. If EVAAS, the “most comprehensive and reliable” VAM available (SAS, 2012a) produces such unreliable results as reported by HISD teachers, I certainly would not trust that any other VAM could further reduce the risk of misclassifying teachers. Further, no matter how much more sophisticated the statistical model becomes, the reality is that human factors and life circumstances inherently impact a student’s ability to learn, and cannot be “controlled for” or deduced from a one-size-fits-all equation. In other words, a

statistical model used to evaluate teachers based on student test data, will likely never have even an acceptable level of reliability, and accordingly will likely always be inappropriate to use to inform consequential decisions.

Validity. HISD teachers provided their opinions about the appropriateness of the student achievement data used for their EVAAS calculations, as related to content-related evidence of validity. The consensus for the majority of teacher respondents was that the data used to calculate their EVAAS scores were representative of the grades, subjects, and students that they actually taught. Approximately 18% of the time, errors were noted. This implies that, although imperfect, the student-to-teacher linking process for EVAAS seems to work well for the majority of teachers in HISD.

The areas of concern came from the teacher participants who were primarily concerned with allocating instructional time among multiple teachers and whether student mobility, in and out of their classrooms, could indeed be controlled with fractional and proportional statistics (see also Corcoran, 2010; Ishii & Rivkin, 2009; Kane & Staiger, 2008; Kennedy, 2010; Nelson, 2011; Papay, 2010; Rothstein, 2009). Although EVAAS can purportedly account for team-teaching dynamics (Sanders & Horn, 1994), it is questionable whether the mathematical proportioning of instruction time without considering the interaction effects of multiple teachers is actually possible (see Amrein-Beardsley & Collins, 2012).

Specifically, co-teachers reported both scenarios of receiving EVAAS scores for students they had never taught, and also not receiving EVAAS scores

for students they had. Multi-grade teachers indicated that they taught more than one grade level per year, but only received EVAAS data for one grade level. Other teachers described situations where students were placed in their classrooms only within weeks of testing, or where students were removed from their classrooms early in the year to attend alternative schools in the district for disciplinary reasons, but these students' scores still showed up on their EVAAS reports. Again, although these examples highlight the imperfections that occurred among a minority of HISD teachers (e.g., 17.5%), the implications of such errors should not be ignored.

Next, in terms of criterion-related evidence of validity, teachers described the relationship between their EVAAS scores and PDAS principal evaluation scores, both of which are considered the main measures of teacher quality in HISD. More than half of the teachers reported that the two evaluations scores did not typically match. The plurality of teachers indicated their PDAS scores were always higher than their EVAAS scores, and that their PDAS scores remained consistent year-to-year while their EVAAS scores fluctuated. Such findings could reflect the subjectivity of the more traditional principal evaluation method, which is believed to lack distinguishability, and largely overestimate the number of effective teachers. This has been dubbed the “Widget” effect (see Weisberg et al., 1999).

Related, it appeared that principals viewed EVAAS as the more objective evaluation score, in that some teachers reported that principals would adjust their PDAS scores (either higher or lower) to reflect their EVAAS scores. This

confounds the criterion-related validity between both measures. Although researchers highly recommend that value-added output correlate with at least one other measure of teacher effectiveness to increase trustworthiness (AERA, APA & NCME, 1999; Baker et al., 2010; Harris, 2011; Hill et al., 2011), such intentional adjustment of scores from one measurement to reflect those of the VAM would completely negate this rationale, yet there is evidence of this occurring elsewhere as well (Garland, 2012; Ravitch, 2012b).

Teachers provided other evidence of teacher quality, that theoretically would complement or counter their EVAAS scores, most often naming awards, recognitions, or leadership roles that contradicted their low EVAAS scores. In fact more than one third of the teachers who responded to this question in the survey instrument reported that they were master or lead teachers, department chairs, or development or academic coaches, having been appointed by peers or principals based on their skillsets and expertise in the very same subject matters in which they were deemed “ineffective” or had received low EVAAS scores.

Validity implications. Although HISD uses two different tools to evaluate teacher effectiveness: EVAAS and PDAS, and although researchers encourage the use of multiple measures to increase validity (AERA, APA & NCME, 1999; Baker et al., 2010; Harris, 2011; Hill et al., 2011), having two that produce conflicting results approximately half of the time, reduces the validity of both measures and sends conflicting messages to teachers.

Further, teachers can only truly assess their work when they have a clear understanding of the targets that their teaching practices are meant to achieve, and

when two indicators of teacher quality produce conflicting results, the targets become even more blurred. This is an important issue to consider as states and districts try to follow recommendations of incorporating multiple measures of teacher quality, recommendations that are currently most popular among academics and researchers (see, for example, Harris, 2011; Hill et al., 2011; Kane & Staiger, 2012; Sass & Harris, 2012); though not enough has been done to this point to determine what level of correlations among multiple measures are appropriate enough to indicate validity. Meanwhile, it seems even two measures of teacher quality cannot be trusted to determine whether a teacher is “effective” or “ineffective,” especially when one appears to influence or trump the value of the other.

Formative use. Data alone cannot increase teacher quality; it is what teachers do with the data that has the potential to make a difference. To investigate whether HISD teachers used the EVAAS data in formative ways, it was necessary to understand how and when they received the EVAAS reports. The teachers indicated discrepancies in the distribution of EVAAS reports to teachers across the district. The majority of teachers received data in the fall for students that had already left their classroom. Formative use then requires teachers to apply what they learned from one group of students to a different group of students, who may not have had the same academic needs as the previous group. The teachers who received EVAAS data for their incoming cohort of students probably would have had the most potential to target individual needs of students. However, those data are technically derived via the prior

teachers' instructional techniques and not their own, and as such are more easily dismissed. Regardless, only 12.1% of the HISD teachers reported receiving student EVAAS data in advance. Other teachers indicated they were only told whether they were "effective" or "ineffective" as rated via EVAAS, oftentimes through the online portal, but these teachers typically did not receive EVAAS data for their students.

Next, teachers reported whether they used EVAAS information to inform their instruction. Teachers who reported using EVAAS data referenced using other data resources in combination to inform their instructional practices; however, almost 60% of the teachers reported that they did not use their EVAAS data for formative purposes whatsoever, and many indicated that they used other data instead, not EVAAS output, to inform their practices. Of the teachers who did report using EVAAS data, the majority called the reports "vague" and "unclear" and the teachers were "not quite sure how to interpret" or use the data to inform their instruction (see also Eckert & Dabrowski, 2010; Harris, 2011).

Other respondents provided statements about looking at the data to provide a general idea about their students, or interestingly enough for ability grouping. This is troublesome in that given the issues with reliability and validity mentioned above, and elsewhere in the literature, meaning these types of consequential decisions should probably not be made. Related, teachers referenced using EVAAS reports to identify the lowest performing kids for remediation or the "bubble kids" on whom they reported focusing their teaching

efforts to try to maximize (or artificially inflate) growth scores (Haladyna, Nolen & Haas, 1991).

As the HISD EVAAS report figures illustrated in Chapters 2 and 3 show, they contain a lot of (albeit confusing) information. And although HISD has offered training sessions, both in-person and online to help teachers understand the reports and how they might use such information to inform their teaching practices, more than one third of teacher participants were unaware of the training sessions. Of the teachers who were aware and had participated in training sessions, the majority found the sessions to be unhelpful.

Beyond these training sessions, though, teacher respondents also reported relying on or looking to their principals for EVAAS information and explanations. Almost half of the teachers indicated that they typically discussed their EVAAS results with their principals, although the other half did not. Among those who did discuss their EVAAS reports with their principals, very few indicated that their principals were able to provide specific information on how they might use the data to improve instruction, however (see also Eckert & Dabrowski, 2010; Harris, 2011). Many teachers believed the “basic discussions” resulted from their principals not understanding EVAAS either, or definitely not understanding EVAAS well enough to explain it to their own teachers. Without principal understanding and buy-in, value-added data are essentially worthless (Kennedy et al., 2012).

Formative use implications. In sum, because teacher respondents indicated that HISD does not have a cohesive district-wide plan for the

distribution and use of EVAAS data, this implies that there are unrealistic assumptions that teachers are using the EVAAS data or that they are aware of the resources available to theoretically help them use the EVAAS data in formative ways. To maximize utility of value-added data, EVAAS reports should be distributed district-wide at the same time. As a result of a cohesive district-wide plan, principals should be provided resources so that they become better equipped at understanding the EVAAS reports. Accordingly, principals might become more able to provide their teachers with specific actions and goals that incorporate the data, develop regular routines to discuss such data, plans and goals with the teachers (Kennedy, et al., 2012), and ensure that all teachers are aware of available training sessions provided by the district. This, however, follows the assumption that the EVAAS data are comprehensible and meaningful, which data from this study contradict. Also recall that not one state representative from the national overview study (Collins & Amrein-Beardsley, under review) could articulate a statewide plan for formative use of the VAM and growth model data, which may indicate the data is not transparent enough to allow for formative use.

Nonetheless, other districts and states looking to implement a VAM should realize that failure to develop a cohesive, unified plan for data dissemination, training, and regular discussions that involve VAM output, will result in a lack of data usage. Principals are fundamental in such plans, particularly as the instructional leaders of their schools. As such, they must not only be knowledgeable about the VAM, but informed of its fine intricacies and related literature base; that is, the academic literature and not just the literature

base advanced by the VAM corporations sponsoring the VAM. Accordingly, principals must be supportive of teachers and encourage the use of these and other data to not only inform their practices, but also question, for example, when things do not make sense. This would increase teachers' and administrators' capacities to become critical consumers. Formative use is the culmination of VAMs, and many, including policymakers, assume that simply enacting legislation which requires states and districts to use such models for summative purposes will simultaneously result in greater levels of data use. However, no states currently have policies or even state-wide plans for using value-added data for formative purposes (Collins & Amrein-Beardsley, under review).

Intended consequences. The large majority of the teachers in this sample strongly disagreed with EVAAS marketing claims and statements. This provides solid evidence that the majority of the teacher respondents do not believe that the EVAAS works in the ways in which both Dr. Sanders and SAS have advertised, to not only HISD at the rate of \$500,000 per year (Amrein-Beardsley & Collins, 2012), but to many other states and districts across the country. Overwhelmingly, teacher respondents reported not believing that the EVAAS model has benefitted much of anything (see, again, each statement as advertised in Table 7 with levels of disagreement).

Intended consequences implications. This signifies that other districts and states need to be, again, critical consumers, and ask for preferentially peer-reviewed evidence to provide accurate, unbiased, and research-based insight into what VAMs look like in practice. It is one thing to judge a book by its cover, or to

read the foreword written by an author's friend, but another completely to read the Consumer Reports, from those who have used the product. In this case that means looking beyond the proprietary company's literature and research on the VAM and gathering feedback from teachers – the “consumers” of VAMs.

Unintended consequences. Throughout teachers' reported experiences and perceptions about EVAAS within HISD, several unintended consequences were also uncovered. As mentioned throughout, teachers repeatedly identified specific groups of students (e.g., gifted, ELL, transition, special education) that typically demonstrated little to no EVAAS growth. Some teachers reported they would (or should) “refuse to teach” such students, if it were not only procedurally but also ethically possible, but given the pressure to obtain high EVAAS scores.

Other teachers described various teaching scenarios such as teaching back-to-back grade levels or switching grade levels which negatively impacted their EVAAS scores. Such reports contradict Dr. Sanders' claim that a teacher in one environment is equally as effective in another (LeClaire, 2011). Also a result of the pressure placed on EVAAS scores, teachers admitted that they can “drill and kill,” teach to the test, or even cheat to effectively, although again artificially (Haladyna et al., 1991), raise their EVAAS scores. Similarly, teachers were able to point out specific grade levels, subjects, and types of students they would both avoid and select if their sole purpose was to obtain the highest EVAAS score. This not only highlighted the fact that teachers believe the EVAAS model produces bias results, but it also demonstrated that teachers believed it can be manipulated or influenced by various criteria and characteristics of the students

assigned to their classrooms (see also Braun, 2005; Hill et al., 2011; Kupermintz, 2003; Rothstein, 2010).

Likewise, teachers explained how EVAAS has created a sense of competition among teachers and has distorted collaboration, for example, when teachers realize that their efforts will go unrecognized and unrewarded, particularly if their actions may contribute to another's EVAAS scores. Researchers have implied such competition could occur when VAMs are used for high-stakes consequences, especially monetary compensation (Harris, 2011; Kennedy et al., 2012), but this remains relatively unexplored. Related, teachers reported that the overall focus on EVAAS scores has lowered morale in their schools as teachers feel overworked and underappreciated.

Unintended consequences implications. As the first study to examine what EVAAS looks like in practice from the perspectives and experiences of HISD teachers, many negative, unintended consequences were discovered as a result of EVAAS use, especially given the high-stake consequences attached to EVAAS output by the district. The evidence here should alarm district administrators, as EVAAS appears to be doing more harm than good, and is potentially preventing students from realizing a well-rounded education. There is even evidence that, at least at a hypothetical level, teachers are becoming increasingly discouraged from working with the very student groups that likely benefit from teachers the most.

Cultural consensus. Using cultural consensus theory as my conceptual framework (Romney et al., 1986), this study was designed to include the input of

those at the center of EVAAS use as a means to evaluate teacher effectiveness, the HISD teachers. First, I found that there is in fact no consensus pertaining to the issue of reliability across EVAAS data from year-to-year, as half of the teachers reported reliable data and the other half did not. However, this matches recent reliability discussions among researchers (see also Baker et al., 2010; Corcoran, 2010; EPI, 2010; Newton et al., 2010; NRC, 2010; Papay, 2010; Rothstein 2010). One could argue that the consensus was that inconsistencies do in fact exist, which indicates that the issue of reliability in the EVAAS and other VAMs remains controversial.

Otherwise, there was consensus among teachers throughout the remainder of the study. First, teachers shared the belief that demographic factors such as home life, health, and behavior impact EVAAS scores, which contradicts the claims that EVAAS can account for such factors. On the topic of validity, the consensus was that the process of linking student to teacher data, though imperfect, worked correctly for most teachers. With criterion-related validity, the consensus was that EVAAS scores did not match PDAS evaluation scores. Additionally, the intended and unintended consequences of EVAAS discovered throughout the study significantly reduced consequential validity. In terms of intended consequences, there was consensus among teachers that the EVAAS reports were seen as vague and unusable, and therefore not used to inform their teaching practices. Additionally, through their responses to a list of EVAAS statements, the consensus was that EVAAS was not seen to benefit anything, including professional development resources, student learning opportunities, or

school culture and teacher morale. In terms of unintended consequences, the consensus was the shared belief that student assignment and specific student characteristics (e.g., ELL, transition, special education, gifted) and factors outside of their teaching control and capabilities, unfairly biased teacher EVAAS scores.

Conclusions

EVAAS and other VAMs, by themselves, are sophisticated statistical models that purportedly provide diagnostic information about student academic growth, and represent teachers' value-add. In other words, EVAAS and VAMs are tools. It is what teachers, schools, districts, and states *do* with this information that matters most. However, for the teachers in this study, even with training sessions, the EVAAS data alone were unclear and virtually unusable. For HISD, not only are teachers not using the “product” that costs the district half a million dollars per year, but teachers are aware that EVAAS inputs can be manipulated based on the student makeup of their classroom, and some teachers even confess to teaching to the test and cheating in attempt to increase their EVAAS scores.

The results from this study provide very important information of which not only HISD administrators should be aware, but also any other administrators from districts or states currently using or planning to use a VAM for teacher accountability. Although high-stakes use certainly exacerbates such findings, it is important to consider and understand that unintended consequences will accompany the intended consequences of implementing this, or likely any other VAM. Reminiscent of Campbell's law, the overreliance on value-added assessment data (assumed to have great significance) to make high-stakes

decisions risks contamination of the entire educational process; for students, teachers and administrators (Nichols & Berliner, 2007).

Accordingly, these findings also strongly validate researchers' recommendations to not use value-added data for high-stakes consequences (Eckert & Dabrowski, 2010; EPI, 2010; Harris, 2011). Though given the EVAAS model's vulnerability as expressed by the HISD teachers, I would advise against using value-added data for anything, at this point, especially as the teachers indicated the EVAAS reports do not provide clear, actionable data that could be used to improve practice.

Yet the general public, motivated largely by the federal government and state governments "racing to the top" to abide, appears to have a lot of faith in these models to reform education by eliminating ineffective teachers from the system, and consequently, yet purportedly, lead to higher student achievement. The trend to adopt VAMs appears to be occurring via commands and promised federal dollars instead of implementing such policies in a holistic manner that encourages and values the input and support of teachers, not to mention the research base surrounding such initiatives. But the failure to consider and incorporate the perspectives and realities of teachers will likely result in yet another one of education's "classic swing of the pendulum...the cycle of early enthusiasm, widespread dissemination, subsequent disappointment, and eventual decline" (Slavin, 1989, p. 752).

Recommendations for Further Study

This study focused on the perspectives and realities of teachers, in the district using the “most comprehensive and reliable” VAM on the market (SAS, 2012a) for more high-stakes consequences than any other district or state in the country. Throughout the study, HISD teachers’ voices and experiences provided necessary insight into what EVAAS looks like in practice. Let us listen to them, learn from them, and work with them, before they leave our schools and our children, behind.

What this study did not do, or did not do well, might inform future research studies in this area. For example, a future study, again in Houston, might attempt to decipher whether teachers are able to distinguish between the EVAAS and ASPIRE systems. It seemed to many of the study participants that these were deemed synonyms, and teasing apart whether participants were responding about one or the other was at times indiscernible.

In addition, the high-stakes use of EVAAS within the ASPIRE program, as well as its impact on termination decisions, likely influenced teacher perceptions and experiences, particularly as the participants in this study were in a district using these value-added data in highly consequential ways. This, in itself, set a tenuous scene, where one might expect adversarial attitudes, on principle alone, from the start. As such, a future study might examine teacher experiences with and perceptions of EVAAS in a district not using data for such high-stakes consequences. This might also add to the research base regarding, in particular, VAMs and criterion-related evidence of validity, in that administrators might feel

less pressured to, for example, skew their more subjective supervisor evaluation scores (see Garland, 2012; Ravitch, 2012b) and teachers might feel less tendencies to attempt to game these systems (Amrein-Beardsley et al., 2010). A future study like this might also investigate, if possible, whether things like the random assignment of students to classrooms indeed reduces, or perhaps eliminates, the impact of student bias on value-added scores.

Future work should investigate the impact of VAM and growth model implementation on student achievement. While it is commonly accepted that teachers are the most important factor contributing to students' in-school learning, it remains unclear whether using VAMs for teacher evaluations actually improves student achievement. Evidence from this study suggests the adverse effects of EVAAS may actually be limiting and inhibiting student access to a well-rounded curriculum. Given their current use, VAM and growth models target only teacher accountability, leaving out student-level accountability and therefore students may not realize or care about the significance that their testing data can have on teachers' lives. Perhaps a future study might also investigate the inclusion of a student-level accountability component to the data used for VAMs, although I strongly hesitate advocating for increased or complimentary accountability systems given the paucity of research evidence that such high-stakes accountability works (Nichols & Berliner, 2007).

A future study might also re-examine formative data use among teachers, again as well, as this might look very different in a district using VAM output for high and low stakes consequences. The latter, by its very being, might be more

likely to take a more reasonable and holistic approach, and hence focus more on the formative versus summative aspects of the VAM implemented. Most related to this study, it would be most interesting to examine teachers' perceptions and experiences with EVAAS or another such VAM in a district that has implemented the EVAAS in a manner that encourages open communication, integrates data in regular discussions, goal setting and planning among teachers and principals, and educates all teachers, principals, and administrators on the methodology, and how to incorporate it into everyday teaching practices. This, of course, assumes such a district exists.

These are just some examples of studies that would be of great interest, particularly in districts or schools that might use these data in less consequential ways. All in all, what these studies can do is provide additional insight on what VAMs look like in practice, as experienced by teachers. Such studies, investigating those most impacted by VAMs – the teachers – will provide the most relevant information on whether VAMs can add value to teaching practices and ultimately increase teacher quality.

But from this study, what we know most importantly is that at least in HISD, the “most comprehensive and reliable” VAM on the market (SAS, 2012a), the EVAAS model, is not working as Dr. Sanders and SAS have intended and marketed, and instead, is resulting in negative, unintended consequences that appear to be harming the teachers, and by mere association, the students whom these teachers teach.

In addition, such high-stakes evaluations are contributing to the deterioration of the teaching profession. The way in which teachers are held accountable for student test scores is unjust. Nichols and Berliner (2007) describe:

[Teachers] are clearly not solely responsible for that performance, but teachers and schools are judged as if they were. Do physicians get punished when their patients supersize their food portions and develop diabetes? Do dentists get punished if their patients will not brush after every meal? (p. 151)

The increasing pressure that teachers and administrators are under will not only exhaust the currently employed teachers and administrators, but also deter others from joining the profession. As Nichols and Berliner (2007) explain, if teachers and administrators no longer desire to remain in schools, how can we expect the schools to be good places to send our children? While our K-12 school days may be long over, we have the tremendous opportunity to learn one more important lesson from teachers, and that is, only teachers can truly educate and inform us on value-added use in practice.

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37, 65-75.
- Amrein-Beardsley, A. (2012). Value-added measures in education: The best of the alternatives is simply not good enough [Commentary]. *Teachers College Record*. Retrieved from <http://www.tcrecord.org/content.asp?contentid=16648>
- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives*, 18.
- Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences. *Education Policy Analysis Archives*, 20.
- Assessment and Accountability Comprehensive Center. (2011). *Key considerations when measuring teacher effectiveness: A framework for validating teachers' professional practices* (AACC Report). San Francisco and Los Angeles, CA: Gallagher, Rabinowitz, & Yeagley.
- Babbie, E. (1990). *Survey research methodology* (2nd ed.). Belmont, CA: Wadsworth Publishing Co.
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., F Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from <http://www.epi.org/publications/entry/bp278>
- Berends, M. (2006). Survey methods in educational research. In J. Green, G. Camilli & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 623-640). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Berliner, D., & Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the public attack on America's public schools*. New York: Perseus Books.

- Bernard, H. R., & Ryan, G. W. (2010). *Analyzing qualitative data: Systematic approaches*. Thousand Oaks, CA: Sage.
- Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. *Yearbook of the National Society for the Study of Education, 103*, 20-50.
- Boudon, R. (1974). *Education, opportunity, and social inequality*. New York, NY: John Wiley & Sons.
- Braun, H. I. (2004). *Value-added modeling: What does due diligence require?* Princeton, NJ: Educational Testing Service.
- Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.
- Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness ranking of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center. Retrieved from <http://nepc.colorado.edu/publication/due-diligence>
- Brophy, J. E. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10*, 245-252.
- Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers*. A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington, D.C.: Dirksen Senate Office Building (research inbrief). Retrieved from <http://www.aera.net/Default.aspx?id=12856>
- Carey, K., & Manwaring, R. (2011). *Growth models and accountability: A recipe for remaking ESEA*. Washington, D.C.: Education Sector.
- Center on Organization and Restructuring of Schools. (1995). *Successful school restructuring*. Madison, WI: Newmann & Wehlage.
- Ceperley, P. E., & Reel, K. (1997). The impetus for the Tennessee value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 133-136). Thousand Oaks, CA: Corwin Press, Inc.
- Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. B. (2008). *Fundamentals of statistical reasoning in education* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

- Coleman, J. S. (1966). *Equality of educational opportunity*. U.S. Government Printing Office, Washington, D.C.: National Center for Educational Statistics.
- Collins, C., & Amrein-Beardsley, A. (2012). *Putting growth and value-added models on the map: A national overview*. Manuscript submitted for review.
- 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*. Providence, RI: Annenberg Institute for School Reform. Retrieved from <http://www.annenberginstitute.org/products/Corcoran.php>
- Creative Research Systems. (n.d.). *The survey system*. Retrieved from <http://www.surveysystem.com/sscalc.htm>
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed). Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education.
- Darling-Hammond, L. (1990). Instructional policy into practice: "The power of the bottom over the top." *Educational Evaluation and Policy Analysis*, 12, 339-347.
- Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind.' *Race, Ethnicity and Education*, 10, 245-260.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93, 8-15.
- Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods*. Englewood Cliffs, NJ: Prentice Hall.
- Dey, I. (1993). *Qualitative data analysis: A user friendly guide for social scientists*. New York: Routledge.
- Duncan, O. D., Featherman, D., & Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.

- Eckert, J., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan*, 91, 88-92.
- Economic Policy Institute. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, D.C.: Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, Shepard.
- Fincher, C. (1985). What is value-added education? *Research in Higher Education*, 22, 395-398.
- Gabriel, R., & Lester, J. (2012) *Constructions of value-added measurement and teacher effectiveness in the Los Angeles Times: A discourse analysis of the talk of surrounding measures of teacher effectiveness*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), Vancouver, B.C., Canada.
- Garland, S. (2012). Tennessee teacher evaluation systems have rough road ahead. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/2012/02/07/tennesseeteacher-evaluat_n_1260790.html?page=1
- Gay, L. R. (1996). *Educational research: Competencies for analysis and application* (5th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.
- Gay, L. R., & Airasian, P. (2003). *Educational research: Competencies for analysis and applications*. Columbus, OH: Merrill Prentice Hall.
- Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA: Sociology Press.
- Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.
- Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems*. Brown Center on Education Policy: The Brookings Institution. Retrieved from www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx
- Goe, L. (2008). *Key issue: Using value-added models to identify and support highly effective teachers*. Washington, D.C.: National Comprehensive Center for Teacher Quality.

- Goldhaber, D., & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. National Center for Analysis of Longitudinal Data in Education Research, Working Paper 31.
- Gravic. (2012). *Remark web survey software*. Retrieved at <http://www.gravic.com/remark/websurvey/>
- Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*, 7-22.
- Greene, J. C., Benjamin, L., & Goodyear, L. (2001). The merits of mixing methods in evaluation. *Evaluation, 7*, 25-44.
- Haertel, E. (2011). *Using student test scores to distinguish good teachers from bad*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.
- Haladyna, T., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*, 2-7.
- Hanushek, E. A. (1970). *The value of teachers in teaching*. Santa Monica, CA: Rand Corporation.
- Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*, 280-288.
- Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources, 14*, 351-388.
- Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*, 466-479.
- Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.
- Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science, 10*-41.
- Hill, D. (2000). He's got your number. *Teachers Magazine, 11*, 42-47.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*, 794-831.

- Houston Independent School District. (2010). EVAAS value-added system myth busters. Retrieved from <http://www.houstonisd.org/HISDConnectDS/v/index.jsp?vnextoid=8b9c8f9da410a210VgnVCM10000028147fa6RCRD&vnextchannel=65db2f796138c010VgnVCM10000052147fa6RCRD>
- Houston Independent School District. (2011). *Board of education workshop 2011-12 budget update* [PowerPoint slides]. Retrieved from www.houstonisd.org/.../Home/.../BudgetUpdate_April212011.ppt
- Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, 4, 520-536.
- Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 25(1), 101-136.
- Jenks, C. S. and others. (1972). *Inequality: A reassessment of the effects of family and schooling in America*. New York: Basic Books.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33, 14-26.
- Kane, T. J., & Staiger, D. O. (2008). *Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.
- Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation.
- Kennedy, B. W. (1991). C. R. Henderson: The unfinished legacy. *Journal of Dairy Science*, 74, 4067-4081.
- Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, 39, 591-598.
- Kennedy, K., Peters, M., & Thomas, M. (2012). *How to use value-added analysis to improve student learning: A field guide for school and district leaders*. Thousand Oaks, CA: Sage.
- Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Working Paper No. 2007-03. Nashville, TN: National Center on Performance Initiatives.

- Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique*. Working Paper N. 2009-01. Nashville, TN: National Center of Performance Initiatives.
- Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives*. Washington, DC: National Academy Press.
- Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value-added assessment system. *Educational Evaluation and Policy Analysis*, 25, 287-298.
- LeClaire, B. (2011). Will EVAAS® make Wake schools better? *Raleigh Public Record*. Retrieved from <http://www.raleighpublicrecord.org/news/2011/06/01/will-evaas-make-wake-schools-better-part-ii/>
- Lincoln Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- Maxwell, J. A. (1941). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.
- Mellon, E. (2010). HISD moves ahead on dismissal policy: In the past, teachers were rarely let go over poor performance, data show. *The Houston Chronicle*. Retrieved from <http://www.chron.com/dispatch/story.mpl/metropolitan/6816752.html>
- Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-66.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, 16, 283-301.

- Milanowski, A. T., Kimball, S. M., & White, B. (2004). The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites. CPRE-UW Working Paper TC-04-01. Madison, WI: University of Wisconsin-Madison. Center for Education Research, Consortium for Policy Research in Education.
- Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.
- Morgan, J. G. (2004). *The education improvement act, a progress report*. Nashville, TN: Comptroller of the Treasury, Office of Education Accountability.
- Morse, J. M. (2003). Principals of mixed methods and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 189-208). Thousand Oaks, CA: Sage.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, D.C.: U.S. GPO.
- National Research Council, & National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop*. Report of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability. Washington, D.C: The National Academies Press.
- Nelson, F. H. (2011). *A guide for developing growth models for teacher development and evaluation*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.
- 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*. Retrieved from <http://epaa.asu.edu/ojs/article/view/810>
- Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.
- Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*. Retrieved from <http://epaa.asu.edu/epaa/v14n1/>

- Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, 20. Retrieved from <http://epaa.asu.edu/ojs/article/view/1048>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.
- Otterman, S. (2010). Hurdles emerge in rising effort to rate teachers. *New York Times*. Retrieved from <http://www.nytimes.com/2010/12/27/nyregion/27teachers.html>
- Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48, 163-193. doi:10.3102/0002831210362589
- Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed.). Needham Heights, MA: Allyn & Bacon.
- Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29, 121-129.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Ravitch, D. (2012a). Flunking Arne Duncan. *The New York Review of Books*. Retrieved from <http://www.nybooks.com/blogs/nyrblog/2012/mar/07/flunking-arne-duncan/>
- Ravitch, D. (2012b). No student left untested. *The New York Review of Books*. Retrieved from <http://www.nybooks.com/blogs/nyrblog/2012/feb/21/no-student-left-untested/>
- Romney, A. K., Weller, S. C., & Batcheled, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313-338.
- Rothstein, J. (2009). *Student sorting and bias in value-added estimation: Selection on observables and unobservables*. Cambridge, MA: The National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w14607>
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125, 175-214.

- Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education* 8, 299-311.
- Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12, 247-256.
- Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student achievement*. Knoxville: University of Tennessee Value-added Research and Assessment Center.
- Sanders, W. L., & Wright, S. P. (2008). *A response to Amrein-Beardsley (2008): "Methodological concerns about the Education Value-Added Assessment System."* Retrieved from www.sas.com/govedu/edu/services/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf
- Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS[®] EVAAS[®]*. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf
- SAS. (2011). *Dr. William L. Sanders* [Biographical sketch]. Retrieved from http://www.sas.com/govedu/edu/bio_sanders.html
- SAS. (2012a). SAS[®] EVAAS[®] for K-12: Assess and predict student performance with precision and reliability. Retrieved from <http://www.sas.com/govedu/edu/k12/evaas/index.html>
- SAS. (2012b). *SAS EVAAS for K-12 Fact Sheet*. Retrieved from <http://www.sas.com/resources/factsheet/education-evaas-factsheet.pdf>
- SAS. (2012c). *SAS EVAAS for K-12 Overview Brochure*. Retrieved from http://www.sas.com/resources/product-brief/SAS_EVAAS_for_K-12.pdf
- SAS. (2012d). *SAS EVAAS for K-12 Validation*. Retrieved from <http://www.sas.com/govedu/edu/k12/evaas/index.html#s1=5>
- Sass, T. & Harris, D. (2012). Skills, productivity and the evaluation of teacher performance. W. J. Usery Workplace Research Group Paper No. 2012-3-1. Retrieved from <http://ssrn.com/abstract=2020717> or <http://dx.doi.org/10.2139/ssrn.2020717>

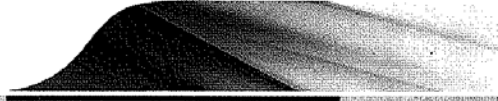
- Saunders, L. (1999). A brief history of educational 'value-added': How did we get to where we are? *School effectiveness and school improvement*, 10, 233-256.
- Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/pubs/20104004/>
- Sewell, W. H., & Hauser, R. M. (1975). *Education, occupation, and earnings: Achievement in the early career*. New York: Academic Press.
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, D.C.: AERA.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, (16)2, 5-24.
- Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan*, 70, 752-758.
- Stake, R.E., & Trumbull, D. (1982). Naturalistic generalizations. *Review Journal of Philosophy and Social Science*, 7, 1-12.
- Stone, J. E. (1999) Value-added assessment: An accountability revolution. In M. Kanstoroom & C. E. Finn (Eds.). *Better teachers, better schools* (pp. 239-250). Washington, D.C.: Fordham Foundation.
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge, MA: Cambridge University Press.
- Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). Thousand Oaks, CA: Sage.
- Tucker, D. P., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- U. S. Department of Education. (2008). *Growth models: Ensuring grade-level proficiency for all students by 2014*. Retrieved from <http://www2.ed.gov/admins/lead/account/growthmodel/proficiency.html>
- U. S. Department of Education. (2009). *Race to the Top program executive summary*. Washington, D. C.: Author.

- U. S. Department of Education. (2012). ESEA flexibility. Retrieved from <http://www.ed.gov/esea/flexibility>
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.
- Wilson, M., Hallman, P. J., Pecheone, R., & Moss, P. (2007, October). *Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's Beginning Educator Support and Training [BEST] Program*. Retrieved from <http://edpolicy.stanford.edu/pages/pubs/pubs.html>
- Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personal Evaluation in Education, 11*, 57-67.
- Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models*. Retrieved from <http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf>

APPENDIX A


ARIZONA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD

APPROVAL



Office of Research Integrity and Assurance

To: Audrey Beardsley FAB

From: Mark Roosa, Chair Soc Beh IRB 

Date: 12/15/2011

Committee Action: Exemption Granted

IRB Action Date: 12/15/2011

IRB Protocol #: 1112007185

Study Title: Houston Independent School District Education Value Added Assessment System

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(1) (2) .

You should retain a copy of this letter for your records.

APPENDIX B

HOUSTON INDEPENDENT SCHOOL DISTRICT RESEARCH APPROVAL

November 10, 2011

Clarín Collins
Principal Investigator and Graduate Student
Mary Lou Fulton Teachers College
Arizona State University
P.O. Box 37100
Phoenix, AZ 85039

Dear Ms. Collins:

The Houston Independent School District (HISD) has granted you the permission to conduct research, as outlined in your proposal titled “The Effects of the Education Value-Added Assessment System (EVAAS) on Teaching Practices and the Profession. The purpose of this research is to understand whether the benefits listed by EVAAS are realities for the teachers impacted by its methodology by studying how EVAAS works in practice. This study will be based on a one-time voluntary survey of HISD teachers to be conducted within a two-week period during the 2011-2012 school year. This study is being conducted jointly by Arizona State University (ASU) and the Houston Federation of Teachers (HFT). The target date for submission of the final report to HISD is May 2012.

Permission to conduct the study in HISD is contingent on your meeting the following conditions:

- The proposed study population will only include teachers of core foundation subjects in 2010-2011 (mathematics, reading/ELA, language, science, social studies) in grades four through eight who have received an EVAAS teacher-level report.
- The study will obtain information from a one-time voluntary survey of a sample of this population.
- The survey will require less than 30 minutes of a teacher’s time and is to be completed outside of school hours.
- The researchers are responsible for identification of the survey population and all data collected.
- This study does not interfere with the District’s instructional/testing program.
- The researcher must follow the guidelines of HISD and the ASU Institutional Review Board (IRB) regarding the protection of human

subjects and confidentiality of data. The HISD signed letter of agreement must be submitted prior to initiating the study.

- While the organization is responsible for oversight of the study, the HISD Department of Research and Accountability may monitor the study to ensure compliance to ethical conduct guidelines established by the Department of Health and Human Services, Office for Human Research Protection (OHRP) as well as the disclosure of student records outlined in Family Educational Rights and Privacy Act (FERPA).
- Data will only be reported in statistical summaries that preclude the identification of any school or teacher participating in the study.
- To eliminate potential risks to study participants, the reporting of proposed changes in research activities must be promptly submitted to the HISD Department of Research and Accountability for approval prior to implementing changes. Non compliance to this guideline could impact the approval of future research studies in HISD.
- The final report must be submitted to the HISD Department of Research and Accountability within 30 days of completion.
- HISD will have complete access to the full data collected by ASU under this research project.

Any changes or modifications to the current proposal must be submitted to the Department of Research and Accountability for approval. Should you need additional information or have any questions concerning the process, please contact me at (713) 556-6700.

Sincerely,
Carla Stevens,
Assistant Superintendent
Research and Accountability

CS: dh
cc: Michele Pola
Alicia Thomas
Chief School Officers
School Improvement Officers
Principals
Julie Baker
Ann Best
Arnold Viramontes
Don Hilber

APPENDIX C
SURVEY PROTOCOL

Verification Questions

1. Are you currently employed by the Houston Independent School District (HISD)?
 - a. Yes
 - b. No

2. How many years have you taught in HISD?
 - c. This is my first year teaching
 - d. 1-3
 - e. 4-5
 - f. 6-10
 - g. 11-15
 - h. 16-20
 - i. 21+

3. How many years have you taught in total?
 - j. This is my first year teaching
 - k. 1-3
 - l. 4-5
 - m. 6-10
 - n. 11-15
 - o. 16-20
 - p. 21+

4. How many years have you received individual Education Value-Added Assessment System (EVAAS) scores? (not school/campus-wide scores)
 - q. 0
 - r. 1
 - s. 2
 - t. 3
 - u. 4
 - v. 5

5. From what type of institution did you receive your teaching certification?
 - w. Public university – in Texas
 - x. Public university – out of Texas
 - y. Private university – in Texas
 - z. Private university – out of Texas
 - aa. HISD certification program
 - bb. Alternative certification program
 - cc. Teach for America
 - dd. Texas Teaching Fellows
 - ee. Other – please specify

6. Including this year, what grade levels have you taught in HISD? (select all that apply)
 - ff. Pre-K
 - gg. Kindergarten
 - hh. 1
 - ii. 2
 - jj. 3
 - kk. 4
 - ll. 5
 - mm. 6
 - nn. 7
 - oo. 8
 - pp. 9
 - qq. 10
 - rr. 11
 - ss. 12

tt. Multi-grade

7. Including this year, what subject areas have you taught in HISD? (select all that apply)

- uu. Mathematics
- vv. Social Studies/History
- ww. Reading/English, Language Arts
- xx. Science
- yy. Music
- zz. Art
- aaa. ESL/Bilingual Education
- bbb. Special Education
- ccc. Test Preparation
- ddd. Physical Education
- eee. Other – please specify

8a. How would you classify the socioeconomic status of students you typically teach in HISD in terms of their needs?

- a. Very high needs
- b. High needs
- c. Average
- d. Low needs
- e. Very low needs
- f. Not applicable

8b. How would you classify the academic status of the students you typically teach in HISD in terms of their needs?

- a. Very high needs
- b. High needs
- c. Average
- d. Low needs
- e. Very low needs
- f. Not applicable

9. Please list your employer organizations (select all that apply):

- a. Congress of Houston Teachers
- b. ATPE (Assoc. of Texas Professional Educators)
- c. HFT (Houston Federation of Teachers)
- d. TCTA (Texas Classroom Teachers Assoc.)
- e. TSTA (Texas State Teachers Assoc.)
- f. Other – please specify

10. What is your gender?

- a. Male
- b. Female

11. What is your identified race?

- a. African American/Black
- b. Asian
- c. Hispanic/Latino(a)
- d. Native American/Indian
- e. Caucasian/White
- f. Two or more races
- g. Other

12. In what year were you born?

(Dropdown menu)

Reliability Questions

13. If you have received more than one year of EVAAS scores, have your scores been consistent over time?

- a. Yes

- b. No, please explain
 - c. Not applicable
14. If you currently teach or have taught more than one grade level, have your scores been consistent across grade levels?
- a. Yes
 - b. No, please explain
 - c. Not applicable
15. If you currently teach or have taught more than one subject level, have your scores been consistent across subject areas?
- a. Yes
 - b. No, please explain
 - c. Not applicable
16. If you currently teach or have taught different types of students (i.e., varied proportions of ELL, gifted, special ed., low/high income), have your scores been consistent regardless of the students you have taught?
- a. Yes
 - b. No, please explain
 - c. Not applicable
17. If there is anything else you would like to add regarding the questions above, please do so here:

Validity Questions

18. Have you ever been evaluated using EVAAS for a grade level for which you were not the teacher of record?
- a. Yes, please explain
 - b. No
 - c. Not applicable
19. Have you ever been evaluated using EVAAS for a subject area for which you were not the teacher of record?
- a. Yes, please explain
 - b. No
 - c. Not applicable
20. Have you ever been evaluated using EVAAS for a group of students for which you were not the teacher of record?
- a. Yes, please explain
 - b. No
 - c. Not applicable
21. Do your EVAAS scores typically match your principal/supervisor observation/evaluation scores?
- a. Yes
 - b. No, please explain
22. Are there any recommendations, awards, student/parent feedback, peer or mentor evaluations that support or contradict your EVAAS scores?
- a. Yes, please explain
 - b. No
23. If there is anything else you would like to add regarding the questions above, please do so here:

Formative Uses & Consequences

24. When do you typically receive EVAAS reports for the students you teach?
- a. Prior to them entering your classroom
 - b. Summer – after students have left your classroom
 - c. Fall – when students are in the next grade level
 - d. You do not typically receive the EVAAS scores for your students
 - e. You have never received the EVAAS scores for your students
 - f. Other, please specify

- 25a. If you have received EVAAS reports for your students, have you used their EVAAS reports to inform your instruction?
- Yes, please explain
 - No
- 25b. With regard to EVAAS data usage, which of the following scenarios describe your situation (check all that apply):
- You use EVAAS data to inform your classroom practices
 - You do not typically use EVAAS data to inform practices
 - You use other resources (not EVAAS data) to inform practices
26. Are you aware of EVAAS training sessions that are available to help you understand the model and reports?
- Yes
 - No
27. Are EVAAS trainings mandatory or optional?
- Mandatory
 - Optional
 - You are not aware of such trainings
28. How many in-person sessions have you attended to better understand EVAAS, your EVAAS scores, how to use your EVAAS scores, etc.?
- 1
 - 2
 - 3
 - 4
 - 5 or more
29. How many online trainings have you attended to better understand EVAAS, your EVAAS scores, how to use your EVAAS scores, etc.?
- 1
 - 2
 - 3
 - 4
 - 5 or more
30. Did you find the EVAAS training sessions helpful?
- Yes
 - No
 - Not applicable
31. Does your principal/supervisor typically reflect on your EVAAS report to improve your instruction?
- Yes
 - No
 - Not applicable
32. To what extent do you typically reflect on your EVAAS report to improve your instruction? Please explain
33. If there is anything else you would like to add regarding the questions above, please do so here:

Overall Questions

To what extent do you agree with the following statements:

- Strongly agree
 - Agree
 - Neither agree or disagree
 - Disagree
 - Strongly disagree
 - Not applicable
- EVAAS reports are simple to use
 - EVAAS helps create professional goals
 - EVAAS helps improve instruction
 - EVAAS ensures growth opportunities for students

5. EVAAS ensures growth opportunities for very low achieving students
6. EVAAS ensures growth opportunities for very high achieving students
7. EVAAS helps increase student learning
8. EVAAS helps you become a more effective teacher
9. EVAAS will validly identify and help to remove ineffective teachers
10. EVAAS will identify excellence in teaching or leadership
11. EVAAS will provide incentives for good practices
12. EVAAS will enhance the school environment
13. EVAAS will enhance working conditions
14. Overall, the EVAAS is beneficial to me as a teacher
15. Overall, the EVAAS is beneficial to my school
16. Overall, the EVAAS is beneficial to the district

34. If your sole purpose as a teacher was to gain the highest EVAAS score, what/who would you select to teach and why?
35. If there is anything else you would like to add regarding the questions above, please do so here:
36. If there is anything else you would like to add overall, please do so here:

APPENDIX D

INTRODUCTION AND REMINDER EMAILS TO HOUSTON

INDEPENDENT SCHOOL DISTRICT TEACHERS

Dear Participant,

I am inviting you to participate in an independent survey research study designed and developed by researchers at Arizona State University (and approved by HISD) to investigate how the data derived via the Education Value-Added Assessment System (EVAAS) has impacted you and your instructional practices.

How this EVAAS survey is different than others you have taken? This is the first study to use the words and experiences of teachers, those most impacted by value-added metrics and models, to gain insight regarding what value-added "looks like" at the classroom level.

Because HISD's use of the EVAAS is at the center of national dialogue about value-added, the opportunity exists to inform these discussions at multiple levels. While this study has been approved by HISD, it was not created or organized by HISD.

This is being conducted by an ASU research team interested in sharing objective findings with the HISD community. This is being done first for local purposes and second to inform national thought and policy.

Your responses will forever remain anonymous and confidential. Honesty is the priority here, and in no way can respondents be identified as per Arizona State University's Institutional Review Board protocols and procedures (IRB # 112007185).

Make your voice count!! Click on the link below to participate in this survey research study. Participation should take approximately 15 minutes.

Note: As you are not to complete this survey during instructional time, feel free to forward this email to your personal email account and complete the survey during non-instructional hours.

Click here to begin: XXXXXXXXXXXXX

Thank you in advance, and if you have any additional questions or concerns please feel free to send me an email at clarin.collins@asu.edu.

Sincerely,

Clarín Collins
PhD Candidate, Educational Leadership & Policy Students
Mary Lou Fulton Teachers College
Arizona State University
Dear Participant,

I am sending a reminder to encourage your participation in a study of the impacts EVAAS has on your teaching practices. Your input has significant potential to inform the nation on what EVAAS looks like in practice. Every response counts – please take 10-15 minutes to share your experience.

Click on the link below to get started, or forward this to your personal email to complete later.

XXXXXXXXXX

Why YOU should participate:

- This is the first EVAAS study in the nation to examine the realities of EVAAS from teachers' experiences
- Your responses are anonymous and confidential
- This study was designed and developed by researchers at Arizona State University, and approved by HISD

Thank you to those of you who have already participated! Please contact me with any questions.

Clarín Collins
PhD Candidate, Educational Leadership & Policy Students
Mary Lou Fulton Teachers College
Arizona State University

APPENDIX E

LIKERT-SCALE TABLE WITH PARTICIPANT RESPONSE PER ITEM

Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Total
1	361 (41.7)	221 (25.5)	146 (16.9)	122 (14.1)	16 (1.8)	<i>N</i> = 866 (100%)
2	346 (39.8)	165 (19.0)	162 (18.6)	173 (19.9)	24 (2.8)	<i>N</i> = 870 (100%)
3	343 (39.7)	173 (20.0)	174 (20.1)	148 (17.1)	26 (3.0)	<i>N</i> = 864 (100%)
4	357 (40.9)	195 (22.3)	182 (20.8)	121 (13.9)	18 (2.1)	<i>N</i> = 873 (100%)
5	362 (41.4)	188 (21.5)	181 (20.7)	122 (13.9)	22 (2.5)	<i>N</i> = 875 (100%)
6	383 (44.0)	185 (21.3)	187 (21.5)	96 (11.0)	19 (2.2)	<i>N</i> = 870 (100%)
7	359 (41.4)	191 (22.0)	184 (21.2)	114 (13.1)	20 (2.3)	<i>N</i> = 868 (100%)
8	383 (44.1)	180 (20.7)	156 (18.0)	123 (14.2)	27 (3.1)	<i>N</i> = 869 (100%)
9	435 (51.2)	184 (21.7)	151 (17.8)	52 (6.1)	27 (3.2)	<i>N</i> = 849 (100%)
10	399 (47.0)	181 (21.3)	166 (19.6)	74 (8.7)	29 (3.4)	<i>N</i> = 849 (100%)
11	365 (42.4)	165 (19.2)	174 (20.2)	117 (13.6)	39 (4.5)	<i>N</i> = 860 (100%)
12	455 (54.0)	156 (18.5)	147 (17.5)	62 (7.4)	22 (2.6)	<i>N</i> = 842 (100%)
13	484 (57.5)	153 (18.2)	143 (17.0)	47 (5.6)	15 (1.8)	<i>N</i> = 842 (100%)
14	415 (48.4)	148 (17.2)	139 (16.2)	124 (14.5)	32 (3.7)	<i>N</i> = 858 (100%)
15	399 (46.7)	144 (16.8)	165 (19.3)	121 (14.2)	26 (3.0)	<i>N</i> = 855 (100%)
16	407 (48.1)	130 (15.3)	172 (20.3)	109 (12.9)	29 (3.4)	<i>N</i> = 847 (100%)

Note. Responses are presented as raw numbers, with respective valid proportions of the total in parentheses.

APPENDIX F

CHI-SQUARE ANALYSES RESULTS FOR LIKERT-SCALE ITEMS

Statement and Chi-square result	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Total
2: EVAAS helps create professional goals; $\chi^2 = (4, N = 870) = 2.11, p = .715$						
Non-HFT	102 (11.7)	52 (6.0)	45 (5.2)	58 (6.7)	9 (1.0)	$n = 266$ (30.6%)
HFT	244 (28.0)	113 (13.0)	117 (13.4)	115 (13.2)	15 (1.7)	$n = 604$ (69.5%)
Total	346 (39.8)	165 (19.0)	162 (18.6)	173 (19.9)	24 (2.8)	$N = 870$ (100%)
3: EVAAS helps improve instruction; $\chi^2 = (4, N = 864) = 2.35, p = .672$						
Non-HFT	104 (12.0)	49 (5.7)	52 (6.0)	53 (6.1)	8 (1.0)	$n = 266$ (30.8%)
HFT	239 (27.7)	124 (14.4)	122 (14.1)	95 (11.0)	18 (2.1)	$n = 598$ (69.2%)
Total	346 (39.7)	173 (20.0)	174 (20.1)	148 (17.1)	26 (3.0)	$N = 864$ (100%)
4: EVAAS ensures growth opportunities for students; $\chi^2 = (4, N = 873) = 3.11, p = .539$						
Non-HFT	104 (11.9)	60 (6.9)	58 (6.6)	42 (4.8)	3 (0.3)	$n = 267$ (30.6%)
HFT	253 (29.0)	135 (15.5)	124 (14.2)	79 (9.0)	15 (1.7)	$n = 606$ (69.4%)
Total	357 (40.9)	195 (22.3)	182 (20.8)	121 (13.9)	18 (2.1)	$N = 873$ (100%)
5: EVAAS ensures growth opportunities for very low achieving students; $\chi^2 = (4, N = 875) = 7.20, p = .126$						
Non-HFT	103 (11.8)	52 (5.9)	62 (7.1)	46 (5.3)	4 (0.5)	$n = 267$ (30.5%)
HFT	259 (29.6)	136 (15.5)	119 (13.6)	76 (8.7)	18 (2.1)	$n = 608$ (69.5%)
Total	362 (41.1)	188 (21.5)	181 (20.7)	122 (13.9)	22 (2.5)	$N = 875$ (100%)
6: EVAAS ensures growth opportunities for very high achieving students; $\chi^2 = (4, N = 870) = 3.96, p = .412$						
Non-HFT	122 (14.0)	48 (5.5)	61 (7.0)	33 (3.8)	4 (0.5)	$n = 268$ (30.8%)
HFT	261 (30.0)	137 (15.7)	126 (14.5)	63 (7.2)	15 (1.7)	$n = 602$ (69.2%)
Total	383 (44.0)	185 (21.3)	187 (21.5)	96 (11.0)	19 (2.2)	$N = 870$ (100%)
7: EVAAS helps increase student learning; $\chi^2 = (4, N = 868) = 7.16, p = .128$						
Non-HFT	105 (12.1)	51 (5.9)	59 (6.8)	45 (5.2)	4 (0.5)	$n = 264$ (30.4%)
HFT	254 (29.3)	140 (16.1)	125 (14.4)	69 (7.9)	16 (1.8)	$n = 604$ (69.6%)
Total	359 (41.1)	191 (22.0)	184 (21.2)	114 (13.1)	20 (2.3)	$N = 868$ (100%)
8: EVAAS helps you become a more effective teacher; $\chi^2 = (4, N = 869) = 3.70, p = .448$						
Non-HFT	108 (12.4)	58 (6.7)	46 (5.3)	45 (5.2)	7 (0.8)	$n = 264$ (30.4%)
HFT	275 (31.6)	122 (14.0)	110 (12.7)	78 (9.0)	20 (2.3)	$n = 605$ (69.6%)
Total	383 (44.1)	180 (20.7)	156 (18.0)	123 (14.2)	27 (3.1)	$N = 869$ (100%)
9: EVAAS will validly identify and help to remove ineffective teachers; $\chi^2 = (4, N = 849) = 4.10, p = .393$						
Non-HFT	124 (14.6)	66 (7.8)	46 (5.4)	19 (2.2)	8 (0.9)	$n = 263$ (31.0%)

HFT	311	118	105	33	19	<i>n</i> = 586
	(36.6)	(13.9)	(12.4)	(3.9)	(2.2)	(69.0%)
Total	435	184	151	52	27	<i>N</i> = 849
	(51.2)	(21.7)	(17.8)	(6.1)	(3.2)	(100%)
10: EVAAS will identify excellence in teaching or leadership; $\chi^2 = (4, N = 849) = 2.58, p = .631$						
Non-HFT	118	64	49	25	8	<i>n</i> = 264
	(13.9)	(7.5)	(5.8)	(2.9)	(0.9)	(31.1%)
HFT	281	117	117	49	21	<i>n</i> = 585
	(33.1)	(13.8)	(13.8)	(5.8)	(2.5)	(68.9%)
Total	399	181	166	74	29	<i>N</i> = 849
	(47.0)	(21.3)	(19.6)	(8.7)	(3.4)	(100%)
11: EVAAS will provide incentives for good practice; $\chi^2 = (4, N = 860) = 3.23, p = .520$						
Non-HFT	105	54	60	31	13	<i>n</i> = 263
	(12.2)	(6.3)	(7.0)	(3.6)	(1.5)	(30.6%)
HFT	260	111	114	86	26	<i>n</i> = 597
	(30.2)	(12.9)	(13.3)	(10.0)	(3.0)	(69.4%)
Total	365	165	174	117	39	<i>N</i> = 860
	(42.4)	(19.2)	(20.2)	(13.6)	(4.5)	(100%)
12: EVAAS will enhance the school environment; $\chi^2 = (4, N = 842) = 0.84, p = .933$						
Non-HFT	138	49	46	22	6	<i>n</i> = 261
	(16.4)	(5.8)	(5.5)	(2.6)	(0.7)	(31.0%)
HFT	317	107	101	40	16	<i>n</i> = 581
	(37.6)	(12.7)	(12.0)	(4.8)	(1.9)	(69.0%)
Total	455	156	147	62	22	<i>N</i> = 842
	(54.0)	(18.5)	(17.5)	(7.4)	(2.6)	(100%)
13: EVAAS will enhance working conditions; $\chi^2 = (4, N = 842) = 2.57, p = .632$						
Non-HFT	145	47	51	15	3	<i>n</i> = 261
	(17.2)	(5.6)	(6.1)	(1.8)	(0.4)	(31.0%)
HFT	339	106	92	32	12	<i>n</i> = 581
	(40.3)	(12.5)	(10.9)	(3.8)	(1.4)	(69.0%)
Total	484	153	143	47	15	<i>N</i> = 842
	(57.5)	(18.2)	(17.0)	(5.6)	(1.8)	(100%)
14: Overall, the EVAAS is beneficial to my school; $\chi^2 = (4, N = 858) = 7.93, p = .094$						
Non-HFT	123	37	52	44	7	<i>n</i> = 263
	(14.3)	(4.3)	(6.1)	(5.1)	(0.8)	(30.7%)
HFT	292	111	87	80	25	<i>n</i> = 595
	(34.0)	(12.9)	(10.1)	(9.3)	(2.9)	(69.3%)
Total	415	148	139	124	32	<i>N</i> = 858
	(48.4)	(17.2)	(16.2)	(14.5)	(3.7)	(100%)
15: Overall, the EVAAS is beneficial to my school; $\chi^2 = (4, N = 855) = 5.162, p = .271$						
Non-HFT	112	46	50	46	6	<i>n</i> = 260
	(13.1)	(5.4)	(5.8)	(5.4)	(0.7)	(30.4%)
HFT	287	98	115	75	20	<i>n</i> = 595
	(33.6)	(11.5)	(13.5)	(8.8)	(2.3)	(69.6%)
Total	399	144	165	121	26	<i>N</i> = 855
	(46.7)	(16.8)	(19.3)	(14.2)	(3.0)	(100%)
16: Overall, the EVAAS is beneficial to the district; $\chi^2 = (4, N = 847) = 5.96, p = .202$						
Non-HFT	115	39	56	42	6	<i>n</i> = 258
	(13.6)	(4.6)	(6.6)	(5.0)	(0.7)	(30.5%)
HFT	292	91	116	67	23	<i>n</i> = 589
	(34.5)	(10.7)	(13.7)	(7.9)	(2.7)	(69.5%)
Total	407	130	172	109	29	<i>N</i> = 847
	(48.1)	(15.3)	(20.3)	(12.9)	(3.4)	(100%)

APPENDIX G
CHI-SQUARE ANALYSES RESULTS FOR ALL OTHER CATEGORICAL
ITEMS

Statement and Chi-square result	N/A	Yes	No	Total
13: If you have received more than one year of EVAAS scores, have your scores been consistent over time? $\chi^2 = (2, N = 874) = 3.589, p = .166$				
Non-HFT	33 (3.8)	101 (11.6)	134 (15.3)	268 (30.7%)
HFT	66 (7.6)	270 (30.9)	270 (30.9)	606 (69.3%)
Total	99 (11.3)	371 (42.4)	404 (46.2)	874 (100%)
14: If you currently teacher or have taught more than one grade level, have your EVAAS scores been consistent across grade levels? $\chi^2 = (2, N = 873) = 2.818, p = .244$				
Non-HFT	107 (12.3)	82 (9.4)	78 (8.9)	267 (30.6%)
HFT	207 (23.7)	205 (23.5)	194 (22.2)	606 (69.4%)
Total	314 (36.0)	287 (32.9)	272 (31.2)	873 (100%)
15: If you currently teach or have taught more than one subject area, have your EVAAS scores been consistent across subject areas? $\chi^2 = (2, N = 867) = 4.251, p = .119$				
Non-HFT	98 (11.3)	76 (8.8)	90 (10.4)	264 (30.4%)
HFT	192 (22.1)	215 (24.8)	196 (22.6)	603 (69.6%)
Total	290 (33.4)	291 (33.6)	286 (33.0)	867 (100%)
16: If you currently teach or have taught different types of students (i.e., varied proportions of ELL, gifted, special ed, low/high income), have your EVAAS scores been consistent regardless of the students you taught? $\chi^2 = (2, N = 877) = 1.448, p = .485$				
Non-HFT	55 (6.3)	95 (10.8)	117 (13.3)	267 (30.4%)
HFT	112 (12.8)	242 (27.6)	256 (29.2)	610 (69.6%)
Total	167 (19.0)	337 (38.4)	373 (42.5)	877 (100%)
18: Have you ever been evaluated using the EVAAS for a grade level for which you were not the teacher of record? $\chi^2 = (2, N = 875) = 0.840, p = .657$				
Non-HFT	10 (1.1)	21 (2.4)	237 (27.1)	268 (30.6%)
HFT	24 (2.7)	59 (6.7)	524 (59.9)	607 (69.4%)
Total	34 (3.9)	80 (9.1)	761 (87.0)	875 (100%)
19: Have you ever been evaluated using EVAAS for a subject area for which you were not the teacher of record? $\chi^2 = (2, N = 874) = 0.218, p = .897$				
Non-HFT	12 (1.4)	24 (2.7)	233 (26.7)	269 (30.8%)
HFT	26 (3.0)	60 (6.9)	519 (59.4)	605 (69.2%)
Total	38 (4.3)	84 (9.6)	752 (86.0)	877 (100%)
20: Have you ever been evaluated using EVAAS for a group of students for which you were not the teacher of record? $\chi^2 = (2, N = 871) = 2.067, p = .356$				
Non-HFT	9 (1.0)	40 (4.6)	218 (25.0)	267 (30.7%)
HFT	25 (2.9)	112 (12.9)	467 (53.6)	604 (69.3%)
Total	34 (3.9)	152 (17.5)	685 (78.8)	871 (100%)

	(3.9)	(17.5)	(78.6)	(100%)
21: Do your EVAAS scores typically match your principal/ supervisor observation/ evaluation scores? $\chi^2 = (1, N = 863) = 3.007, p = .083$				
Non-HFT		124	141	265
		(14.4)	(16.3)	(30.7%)
HFT		242	356	598
		(28.0)	(41.3)	(69.3%)
Total		366	497	863
		(42.4)	(57.6)	(100%)
22: Are there any recommendations, awards, student/ parent feedback, peer mentor evaluations that contradict your EVAAS scores? $\chi^2 = (1, N = 843) = 0.028, p = .866$				
Non-HFT		113	144	257
		(13.4)	(17.1)	(30.5%)
HFT		254	332	586
		(30.1)	(39.4)	(69.5%)
Total		367	476	843
		(43.5)	(56.5)	(100%)
25: If you have received EVAAS reports for your students, have you used their EVAAS reports to inform your instruction? $\chi^2 = (1, N = 815) = 0.027, p = .868$				
Non-HFT		98	143	241
		(12.0)	(17.5)	(29.6%)
HFT		237	337	574
		(29.1)	(41.3)	(70.4%)
Total		335	480	815
		(41.1)	(58.9)	(100%)
26: Are you aware of EVAAS training sessions that are available to help you understand the model and reports? $\chi^2 = (1, N = 870) = 1.373, p = .241$				
Non-HFT		174	91	265
		(20.0)	(10.5)	(30.5%)
HFT		372	233	605
		(42.8)	(26.8)	(69.5%)
Total		546	324	870
		(62.8)	(37.2)	(100%)
27: Are EVAAS trainings mandatory or optional? $\chi^2 = (2, N = 863) = 1.360, p = .507$				
	Mandatory	Optional	Not aware	
Non-HFT	68	128	72	263
	(7.3)	(14.8)	(8.3)	(30.5%)
HFT	136	276	188	600
	(15.8)	(32.0)	(21.8)	(69.5%)
Total	199	404	260	863
	(23.1)	(46.8)	(30.1)	(100%)
30: Did you find the EVAAS trainings helpful? $\chi^2 = (2, N = 864) = 2.259, p = .323$				
Non-HFT	64	75	122	261
	(7.4)	(8.7)	(14.1)	(30.2%)
HFT	178	161	264	603
	(20.6)	(18.6)	(30.6)	(69.8%)
Total	242	236	386	864
	(28.0)	(27.3)	(44.7)	(100%)
31: Does your principal/ supervisor typically discuss your EVAAS results with you? $\chi^2 = (2, N = 868) = 0.212, p = .899$				
Non-HFT	16	127	124	267
	(1.8)	(14.6)	(14.3)	(30.8%)
HFT	33	295	273	601
	(3.8)	(34.0)	(31.5)	(69.2%)
Total	49	422	397	868
	(5.6)	(48.6)	(45.7)	(100%)