education policy analysis archives

A peer-reviewed, independent, open access, multilingual journal

epaa aape

Arizona State University

Volume 20 Number 12 April 30th 2012 ISSN 1068-2341

The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences

Audrey Amrein-Beardsley
Arizona State University

Clarin Collins
Arizona State University

Citation: 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* (12). Retrieved [date], from http://epaa.asu.edu/ojs/article/view/1096

Abstract: The SAS Educational Value-Added Assessment System (SAS® EVAAS®) is the most widely used value-added system in the country. It is also self-proclaimed as "the most robust and reliable" system available, with its greatest benefit to help educators improve their teaching practices. This study critically examined the effects of SAS® EVAAS® as experienced by teachers, in one of the largest, high-needs urban school districts in the nation – the Houston Independent School District (HISD). Using a multiple methods approach, this study critically analyzed retrospective quantitative and qualitative data to better comprehend and understand the evidence collected from four teachers whose contracts were not renewed in the summer of 2011, in part given their low SAS® EVAAS® scores. This study also suggests some intended and unintended effects that seem to be occurring as a

Manuscript received: 03/01/2012
Revisions received: 03/03/2012
Accepted: 03/04/2012

result of SAS® EVAAS® implementation in HISD. In addition to issues with reliability, bias, teacher attribution, and validity, high-stakes use of SAS® EVAAS® in this district seems to be exacerbating unintended effects.
Keywords: value-added models (VAMs); teacher effectiveness; teacher quality; teacher evaluation; accountability.

El SAS Sistema de Evaluación de la Educación de Valor Agregado (SAS® EVAAS®): Algunos efectos intencionales y no intencionales en un sistema escolar urbano de gran tamaño.
Resumen: El SAS Sistema de Evaluación de la Educación de Valor Agregado (SAS® EVAAS®) es el sistema de valor agregado más ampliamente utilizado en el país. También se auto-proclama como el sistema disponible "más robusto y fiable", siendo su mayor beneficio poder ayudar a los educadores a mejorar sus prácticas de enseñanza. Las investigadoras de este estudio examinaron los efectos de SAS® EVAAS®, experimentado por docentes, en uno de los distritos escolares urbanos más grandes y con mayores necesidades educativas de la nación: el Distrito Escolar Independiente de Houston (HISD). A través de evidencias obtenidas con cuatro docentes cuyos contratos no fueron renovados en el verano de 2011, en parte debido a sus bajas puntuaciones en SAS® EVAAS®, se detallan algunos de los efectos deseados y no deseados que parecen estar ocurriendo como resultado de la aplicación SAS® EVAAS® en HISD. Además de los problemas con la fiabilidad, el sesgo, la atribución docente, la validez, el uso de SAS® EVAAS® con consecuencias severas en este distrito parece exacerbar los efectos no intencionales.
Palabras clave: modelos de valor añadido (VAM); eficacia docente; calidad docente; evaluación docente; rendición de cuentas.

O SAS Sistema de Avaliação do Valor Agregado Educativo (SAS® EVAAS®): Alguns efeitos intencionais e não intencionais em um sistema escolar urbano de grande tamanho.
Resumo: O SAS Sistema de Avaliação do Valor Agregado Educativo (SAS® EVAAS®) é o sistema mais utilizado de valor agregado no país. Também é auto-proclamado como o sistema disponível "mais robusto e confiável", sendo seu maior benefício ajudar a os educadores a melhorar suas práticas de ensino. As pesquisadoras deste estudo examinaram os efeitos da SAS® EVAAS®, com professores que tiveram essa experiência em um dos maiores distritos escolares urbanos e com maiores necessidades educacionais da nação: o Houston Independent School District (HISD). Através de evidencias obtidas com quatro professores cujos contratos não foram renovados no verão de 2011, em parte devido à suas baixas pontuações em SAS® EVAAS®, se analisam os efeitos desejados e indesejados que parecem estar a ocorrer como resultado da aplicação SAS® EVAAS® em HISD. Além dos problemas com a fiabilidade, a polarização, a atribuição de ensino, a validade, a utilização de SAS® EVAAS® em avaliações com conseqüências severas neste distrito parece exacerbar os efeitos indesejados.
Palavras-chave: modelos de valor agregado (VAM); eficácia dos professores; qualidade dos professores; avaliação de professores; prestação de contas.

Introduction

Since the implementation of No Child Left Behind (NCLB) in 2002, researchers, econometricians, and statisticians have explored different analytical methods to document students' academic progress over time, specifically to replace Adequate Yearly Progress (AYP) measures. More recently, President Obama's *Race to the Top* competition (2009) encouraged similarly oriented initiatives, contributing over $350 million in federal support (Robelen, 2012)

to be allocated to those states that adopt methods to better measure the "value" a teacher "adds" to student learning from year to year.

In theory, value-added models (VAMs) allow for richer analyses of test score data because students are *simply* followed to assess their learning trajectories from the time they enter a teacher's classroom to the time they leave. In practice, however, these models do not seem to work in many of the ways theorized. For example, it still remains uncertain whether teachers are accurately classified as contributing to differential gains, whether teachers teaching certain types of students (e.g., special education, gifted, and English Language Learners (ELLs)) are fairly assessed, and whether teachers are using value-added output to inform instructional modifications and improvements (Au, 2010; Eckert & Dabrowski, 2010; Haertel, 2011; Hill, Kapitula, & Umland, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Papay, 2010; Rothstein, 2009). In addition, while the implementation and use of VAMs for high-stakes purposes is increasing across the country, there lingers a paucity of research evidence to support the attachment of significant consequences to value-added output (Braun, 2005; Harris, 2011; Ho, Lewis, & Farris 2009; Schochet & Chiang, 2010).

The purpose of this study was to contribute to the existing research base by critcally examining some intended and unintended effects of the largest and most commonly used VAM – the SAS Education Value-Added Assessment System (SAS® EVAAS®) – in the Houston Independent School District (HISD). This district is using value-added data more than any other in the country for high-stakes purposes, expressly for merit awards and to make teacher termination decisions (Corcoran, 2010; Harris, 2011; Mellon, 2010; Otterman, 2010; Papay, 2010). During the summer of 2011, the two researchers examined SAS® EVAAS® data from four teachers whose contracts were not renewed in terms of reliability, bias, teacher attribution, and validity. They examined other intended consequences (e.g., value-added use and data informed change) and unintended consequences (e.g., perverse side effects) as well.

The SAS Education Value-Added Assessment System (SAS® EVAAS®) and the Houston Independent School District (HISD)

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. While Tennessee, North Carolina, Pennsylvania, and Ohio use SAS® EVAAS® statewide, and other states, districts, and schools are using or have plans to implement this model locally, no other school, district, or state uses SAS® EVAAS® for consequential decision-making more than HISD (Harris, 2011; Lowrey, 2012; Sparks, 2011).

In 2007, HISD created the Accelerating Student Progress: Increasing Results & Expectations (ASPIRE) program to recognize and celebrate great teaching as measured by student progress (HISD, 2010). District administrators contracted with the SAS software company to measure this progress via their SAS® EVAAS® system; this 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 SAS® EVAAS® scores to rank order teachers throughout the district and 2) the Professional Development and Appraisal System (PDAS), in which teacher observation data is collected by certified appraisers and used to

evaluate teachers in eight different domains of teacher performance.[1] 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 and 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).

Regardless, with over 20 years of development, the system in use – the SAS® EVAAS® – is the largest, most widely implemented, and most widely used VAM in the country. While there are at least eight entities developing such models (Banchero & Kesmodel, 2011), like the VAM developed by the Value Added Research Center (VARC) in Wisconsin and the growth model developed by Dr. Damian Betebenner (the Student Growth Percentiles (SGP) model), SAS® EVAAS® is "the most comprehensive reporting package of value-added metrics available in the educational market" (SAS, 2012). It is "the most robust and reliable" system available, better than the "other simplistic models found in the market today" (SAS, 2012). It "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, 2012). These claims are not without controversy, however (Amrein-Beardsley, 2008; Sanders & Wright, 2008). Researchers used these assertions to frame this study, in particular in terms of the intended or expected outcomes, as advertised, as well as the unintended outcomes researchers discovered along the way.

Preliminary Evidence

Even though the district reported that the majority of teachers favor the ASPIRE program overall (Harris, 2011), researchers found evidence suggesting that HISD teachers have aversions towards the program's SAS® EVAAS® component (Collins, in progress). In terms of reliability, those receiving merit monies attached to their SAS® EVAAS® output often compare winning the rewards to "winning the lottery," given the random, "chaotic," year-to-year instabilities they see. Such consistencies are also well noted in literature (Baeder, 2010; Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn et al., 2010; Haertel, 2011; Koedel & Betts, 2007; Papay, 2010). 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 SAS® EVAAS® rankings "jump around." Along with the highs come much-appreciated monetary awards, but for what teachers did differently from one year to the next remains unknown.

Teachers who do not receive merit monies attribute the lack of rewards to the types of students they teach and how these students might bias their scores (Collins, in progress; see also Hill et al., 2011; Newton et al., 2010; Rothstein, 2009). Teachers who loop or teach back-to-back grade levels report bonuses for the first year and nothing the next as they "max out" on growth the first year with the same students. Teachers of grades in which ELLs are transitioned into mainstreamed English-only classrooms report being the least likely to demonstrate added value and the most likely to be deemed "ineffective." Teachers of inordinate numbers of special education students express similar concerns (Collins, in progress; see also Hill et al., 2011; Newton et al., 2010; Rothstein, 2009).

[1] 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).

There are also ceiling effects prevalent, whereas teachers of gifted students experience difficulties demonstrating added value (see also Wright et al., 1997).

Almost half (46%) of a sample of HISD teachers who moved to different grade levels reported switching value-added ranks after the move, from "ineffective" to "effective" or vice versa, also across grade levels that were adjacent (Collins, in progress). This is problematic as the SAS® EVAAS® system is purported to measure the teacher effectiveness construct consistently, and validly. Dr. William L. Sanders, the developer of the SAS® EVAAS®, claims that teachers who move from one environment to another, even if radically different, continue to do just as well (LeClaire, 2011).

Furthermore, over half (55%) of a same sample of HISD teachers noted that their SAS® EVAAS® reports did not match their supervisors' observational PDAS scores (Collins, in progress). In addition, some suggest that their supervisors are skewing their observational scores to match their SAS® EVAAS® scores given external pressures to do so (Collins, in progress). Such practices have been shown to occur elsewhere with the Tennessee Value-Added Assessment System (TVAAS) from which the SAS® EVAAS® was derived (Garland, 2012). In New York as well, if teachers have two years of low value-added scores, the teachers are to be rated ineffective overall and terminated, regardless of what other measures (e.g., supervisor evaluation scores) indicate or disclose (Ravitch, 2012). Because these other measures are often perceived as less objective, it seems that measuring teacher effectiveness using value-added output is beginning to trump other indicators capturing what it means to be an effective teacher. This raises major concerns about cogency and power (i.e., evidence of criterion-related validity). Such practices also contradict the field standards developed by the prominent national associations on educational measurement and testing (AERA, APA, & NCME, 2000). These standards note first and foremost 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, 2000).

Ten percent of the same teachers (10%) noted substantive concerns about being evaluated for content they were not teaching, or being held accountable while teaching alongside other teachers teaching the same students the same subjects at the same time (Collins, in progress). SAS® EVAAS® methodologists state they can adequately control for this using fractions and proportional contributions, however (Derringer, 2010; Sanders & Horn, 1994).

Numerous teachers, especially science and social studies teachers teaching non-tested subjects in every grade level, also note issues when norm-referenced tests are used with criterion-referenced tests to determine SAS® EVAAS® growth from year to year. They note concerns about the pretest scores used to calculate value-added coming from different tests than the post-test scores, and vice versa. Additionally, they note concerns about the norm-referenced tests not being linked to state standards. While norm-referenced and criterion-referenced tests can be normed, and this is somewhat common, this still raises issues with content alignment (i.e., evidence of content-related validity).

In terms of formative uses, because SAS® EVAAS® output is often received months after students leave, teachers express that such output makes little sense, and they are learning little about what they did effectively or how they might use SAS® EVAAS® data to improve their own instruction (see also Eckert & Dabrowski, 2010; Harris, 2011). Of the same sample of HISD teachers surveyed, the majority (55%) note that they receive their SAS® EVAAS® reports in the summer or fall after students leave their classrooms. A plurality (40%) also reported they were unaware of HISD-sponsored professional development trainings about how to better understand or use their SAS® EVAAS® data to improve their instruction (Collins, in progress). This is problematic

since SAS® EVAAS®'s principal claimed strength is to provide a "wealth of positive diagnostic information" for formative purposes (Sanders, Wright, Rivers & Leandro, 2009, p. 9).

HISD, SAS® EVAAS®, and Teacher Non-Renewals

In the spring of 2011, HISD did not renew 221 of its teachers' contracts (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 [SAS® EVAAS®] value-added scores." HISD did not respond to researchers' 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 SAS® EVAAS® scores in spring of 2011, however, so it is uncertain how many teachers were actually terminated for these reasons. All that is known is that, according to one of the lead lawyers retained in these teachers' defense (A. Reichek, personal communication, June 8, 2011), a number of HISD teachers' non-renewal letters cited these reasons for termination. According to the Vice President of the Houston Federation of Teachers (HFT), this number was greater than 100 or nearly 50% (Z. Capo, personal communication, April 6, 2012). Researchers 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. Researchers are, however, aware that attaching such high-stakes decisions to VAM output in general is expected "to lead to a flood of litigation challenging teacher dismissals" as "value added modeling as a basis for high stakes decision making is fraught with problems likely to be vetted in the courts" (Baker, 2012). What researchers examined here are four such cases.

Participants

For four of the terminated teachers, the same lead lawyer invited one of the researchers, (the first author, hereafter referred to as the primary researcher) to serve as the expert witness and testify on their behalves. In terms of sampling procedures, the primary researcher did not select the four teachers with any methodological reason or representative sampling approach. Rather, the teachers quasi-selected the researcher via their lawyer. The lawyer retained the primary researcher to testify regarding (1) the SAS® EVAAS® in general, (2) whether SAS® EVAAS® output for each teacher accurately evidenced that the teacher positively or negatively impacted student achievement and growth, and (3) whether the grounds and reasoning on which their contracts were not renewed were justifiable and sound.

The teachers, four female elementary school teachers, were from racial minority backgrounds (three were African American and one was Latina). Their ages ranged from 28-51. They collectively averaged 11.8 years of teaching experience and 7.5 years teaching in HISD. Two were certified via a traditional teacher certification program and the other two were certified via HISD's Alternative Teaching Certificate (ATC) program. All teachers taught core subject areas (reading, language arts, math, social studies, and science) in grades 3-7, and they all taught in different schools under different school administrations.

It should be noted, here, that given the sensitive nature of these teachers' experiences and testimonies, the primary researcher also secured the four teachers' signed permissions to use the data collected for the lawsuit also for presentation and publication purposes. The primary researcher consulted with each participant about confidentiality, her general rights, and in particular her right to opt out of the study or pull her data from study inclusion at any time, and if at any time she felt she was at risk or to be placed at risk in the foreseeable future. The primary researcher also gave each

participant the contact information for the Chair of the Human Subjects Institutional Review Board, through the Arizona State University Office of Research Integrity and Assurance (IRB Study #11108006705).

Multiple-Methods, Case Study Approach

The researchers conducted a case study (Campbell, 1975; Flyvbjerg, 2011; Ragin & Becker, 2000; Thomas, 2011a) using multiple-methods to examine the collective cases of the four units at focus (Gerring, 2004). The cases were similar and separate enough to permit such an analysis, especially as each of the teachers had associated experiences and could serve as comparable instances of the same general phenomenon (Ragin & Becker, 2000). Their practical experiences (Flyvbjerg, 2011) could help others better understand how this value-added system was being used within HISD.

The primary researcher collected retrospective quantitative and qualitative data to better comprehend and understand the four teachers' data that more holistically captured their effectiveness as teachers (Creswell, 2008; Day, Sammons, & Gu, 2008; Greene, Caracelli, & Graham, 1989; Johnson & Onwuegbuzie, 2004). The quantitative documents included each teacher's SAS® EVAAS® scores and supervisor observational scores based on the district's PDAS system, and the qualitative information came from the written comments provided on each teacher's PDAS forms for the same years for which SAS® EVAAS® data were collected and from the in-depth phone interviews the primary researcher also conducted.

Specifically, the primary researcher collected each teacher's SAS® EVAAS® Teacher Value-Added Reports (see Figure 1). The district in contract with SAS provides such reports, alongside SAS® EVAAS® Reports for Teacher Reflection (see also Figure 1), to teachers yearly through an online portal. These reports include a color chart intended to offer teachers a graphical display of how their different students (low, middle, and high performing) progressed in their classrooms as compared to the district average. The reports also include a table to complement the chart and quantify the colors displayed. Resource guides are available to help consumers understand these reports as well (see, for example, SAS, 2007)

As written, these reports are to be used to evaluate how well individual teachers facilitate student achievement on Texas's Assessment of Knowledge and Skills (TAKS and TAKS Accommodated) and Stanford/Aprenda achievement tests that are used in non-TAKS grades and subject areas. These reports are used to compare how well teachers influence student progress as compared to similar teachers within the district. While scores include an individual teacher's normal curve equivalent (NCE) gain, a measure of standard error for confidence, and a district reference gain also expressed as an NCE indicating how the district did compared to the state average each year, the score of interest here was the gain score index. This score compares each teacher to other similar teachers across the district, and this is the score that HISD uses for determining ASPIRE awards.[2]

[2] As per the statistical rules and policies put in place by SAS® EVAAS® comparisons are made based on one standard error. Teachers with a score above 1.0 are deemed as adding value, teachers with a score between 1.0 and -1.0 are deemed as not detectably different (NDD), and teachers with a score below -1.0 are deemed as detracting value, comparatively.

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

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.**

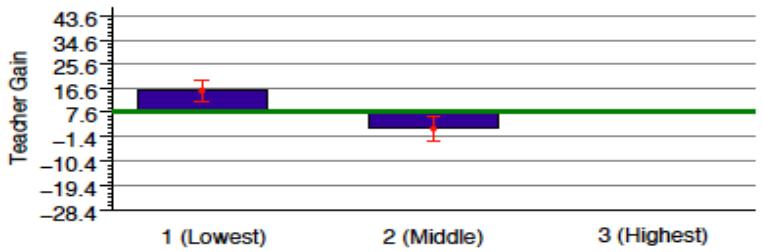
HISD Reference Gain
The **HISD Reference Gain** is the gain made by HISD in this subject and grade. This gain is expressed as an NCE, based on the state population in 2006. A positive gain indicates HISD made more progress than the state average and a negative gain indicates HISD made less progress than 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 1. Teacher C's SAS® EVAAS® Teacher Value-Added Report for 2010 and SAS® EVAAS® Report for Teacher Reflection.

The primary researcher also collected each teacher's PDAS supervisor evaluation scores. She collected their PDAS scores, as they are also valued in HISD's ASPIRE system, for the same years as each teacher's SAS® EVAAS® scores to help contextualize and better understand each teacher's SAS® EVAAS® data. On the PDAS, both numerical scores (i.e., scores marked for each of the eight domains included on the PDAS instrument and overall) and supervisors' written comments (i.e., by domain and overall) were collected for analysis (see a listing of these domains in Sidebar 1; see also PDAS, 2004).

Sidebar 1
The Eight Observational Domains of the Professional Development Appraisal System (PDAS)[3]

I. Active Successful Participation in the Learning Process
II. Learner-Centered Instruction
III. Evaluation and Feedback on Student Progress
IV. Management of Student Discipline, Instructional Strategy, Time, and Materials
V. Professional Communication
VI. Professional Development
VII. Compliance with Policies, Operating Procedures and Requirements
VIII. Improvement of Academic Performance of all Students on the Campus

While it is likely that observational scores are often inflated, and this is in large part why more objective measures of teacher effectiveness like VAMs are adopted and implemented (Jacob & Lefgren, 2007; Harris, 2009, 2011; Ravitch, 2012), it was also important to examine whether in fact observational measures correlated with SAS® EVAAS® output (see also Milanowski, Kimball, & White, 2004). The primary researcher collected this information here because general evidence is lacking to indicate that this measure of teacher value-added is indeed related to at least one other correlated criterion (i.e., evidence of criteria-related validity). The primary researcher also collected other indicators teachers might have had for the same years of analysis, especially about their effectiveness as teachers and to further examine this type of validity.

Finally, the primary researcher collected qualitative data via extensive phone interviews. The researcher spoke with each of the four teachers by phone in the summer of 2011 an average of 2.5 hours per interview. She followed-up with shorter phone calls for verification purposes on occasion. During each phone interview, she first asked each teacher a set of demographic questions (teaching certification, number of total years teaching and teaching in HISD, age, and racial backgrounds). She then asked each teacher to explain their corpus of data for each school year, as aligned with the aforementioned documents. Last, she asked each teacher an additional four, open-ended questions to get at any information that might have been missed and to take a preliminary look at data comprehension, use, and levels of professional support for both. The primary researcher asked each teacher the following:

- Is there anything else you can think of in terms of reasons why your contract is not being renewed (e.g., excessive absenteeism, insubordination, other test scores)?
- Do you understand your SAS® EVAAS® value-added scores?
- Have you received training on how to understand your SAS® EVAAS® reports/scores?
- Have you received professional development as a result of your SAS® EVAAS® scores?

[3] To find out more about these domains, including information about the subscores situated within each domain, see (PDAS, 2004).

Data Analysis

The primary researcher transcribed the interview data and analyzed the transcripts alongside the numerical data, year-by-year, to establish a longitudinal chain of evidence (Yin, 1994). Specifically, the primary researcher analyzed the data by case, and then compared incidents within individuals over time. The researcher developed working assertions across cases, as well, to integrate and develop broader themes (Glaser & Strauss, 1967; Lincoln & Guba, 1985; Patton, 2001).

The teachers involved in the study verified results and findings via a series of member-checks (Guba & Lincoln, 1981). The four teachers read the final report and checked it for accuracy and authenticity, clarified misunderstandings and misconceptions, and verified the overall findings. Researchers also resituated the findings within the literature if they added to specific topics about value-added methods and systems specifically or in general.

It is important to highlight that the experiences of these four teachers should not, however, be used to generalize beyond HISD or to all teachers in HISD. Nevertheless, the researchers are confident that their findings still deliver a strong message and may generalize to the other approximately 100 plus teachers whose 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 [SAS® EVAAS®] value-added scores." Even with a limited, non-representative sample of four, patterns and overall findings may also help others understand this particular value-added system better, via the lived experiences of these teachers in HISD (Feagin & Orum, 1991; Yin 1994).

Results

Teacher A

Teacher A, a university-certified teacher, was an elementary school teacher in HISD since 2000. Illustrated in Table 1 is a summary of Teacher A's SAS® EVAAS® and PDAS scores and ASPIRE bonuses since 2007, the first year of HISD's ASPIRE system.

Table 1
Teacher A's SAS® 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

*Notes: Scores shaded as green indicate that the teacher added value according to SAS® EVAAS® data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (1) Scores with asterisks (*) do 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 gain scores of other teachers across HISD given one standard error; however, the scores are still reported to both the teachers and their supervisors as they are here.

Across all years and subject areas for which Teacher A had SAS® EVAAS® data, she added value to her students' learning (relative to all other HISD teachers) 50% of the time (8/16 of SAS® EVAAS® observations), and detracted value (relative to all other HISD teachers) the other 50% of the time (8/16 of SAS® EVAAS® observations). According to these SAS® 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 [SAS® EVAAS®] value-added scores."

Analyzing Teacher A's SAS® EVAAS® scores alongside her PDAS scores, it is not only visually obvious that there is something peculiar about the relationship between Teacher A's performance on the SAS® EVAAS® and her supervisor evaluation scores, it is also statistically evident. The correlation between Teacher A's SAS® 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 SAS® 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 ($r = -0.42$). Until 2010-2011, Teacher A "exceeded expectations" across every PDAS domain, and her colleagues recognized her as both a "Teacher of the Month" and the "Teacher of the Year" in 2010.

Otherwise, Teacher A was only familiar with SAS® EVAAS® due to the score reports distributed each year and because her colleagues and supervisors used to talk about something called "value-added." Nobody ever explained her SAS® EVAAS® scores to her, and she never fully understood what the numbers meant, how they could impact or "hurt her," or how she could use her SAS® EVAAS® scores to help her improve her own instruction. Additionally, she never received professional development as a result of her value-added scores, although whether she needed professional development to help her improve her value-added scores is questionable.

Teacher B

Teacher B, a career-changer with a bachelor's and master's degree in mathematics, was certified as a math teacher for grades 2-12 via HISD's Alternative Teaching Certificate (ATC) program in 2007. Illustrated in Table 2 is a summary of Teacher B's SAS® EVAAS® and PDAS scores and ASPIRE bonuses scores since 2008.

Table 2
Teacher B's SAS® 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
ASPIREBonus	$1,750	$0	$4,700	n/a

*Notes: Scores shaded as green indicate that the teacher added value according to SAS® EVAAS® data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (1) Scores with asterisks (*) do 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 are still reported to both the teachers and their supervisors as they are here.

Teacher B's relative value-added scores were negative for math for two years, and positive for the most recent year for which she had SAS® EVAAS® data. In her most recent position for which she had SAS® EVAAS® data she seemed to have added value to her students' learning. She taught alongside another math teacher who taught nearly half of her students math an equal amount of time per week all year. Whether she alone demonstrated "a significant lack of student progress attributable to the educator," or "insufficient student academic growth reflected by [SAS® EVAAS®] value-added scores" is debatable. In addition, value-added researchers agree that at least three years of value-added data are needed to make such judgments (Baker, 2012; Brophy, 1973; Cody, McFarland, Moore, & Preston, 2010; Harris, 2011), and even then a 25% risk of misclassification remains (Au, 2010; CCSSO, 2010; Otterman, 2010; Schochet & Chiang, 2010; Shaw & Bovaird, 2011). She did not have three years of consistent data, and her most recent year was demonstrably her best.

Analyzing Teacher B's SAS® EVAAS®scores alongside her PDAS scores, there is a strong relationship between Teacher B's SAS® EVAAS® and supervisor evaluation scores (r = 0.91). The better Teacher B did on the SAS® EVAAS® the better she did in the eyes of her supervisor(s), and vice versa. This yields 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 was monetarily rewarded in a way that made sense; the better she did the more money she received (r = 0.93).

Otherwise, the knowledge that Teacher B had about the SAS® EVAAS® was also sparse. She did not understand how "they" calculated her value-added scores. She would "just see the scores." She also knew that "they" compared her scores "to everybody else's in the district." This teacher did not receive training to understand, or professional development to improve her value-added scores, although whether her most-recent value-added scores were in need of improvement is unclear.

Teacher C

Teacher C graduated with a bachelor's degree in early childhood education in 1999, and received a master's degree in school counseling in 2000. Thereafter, she served as a long-term substitute in HISD until she took a full-time teaching position in HISD, teaching 6th grade in 2003. Illustrated in Table 3 is a summary of Teacher C's SAS® EVAAS® and PDAS scores and ASPIRE bonuses since 2007.

Table 3
Teacher C's SAS® 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

*Notes: Scores shaded as green indicate that the teacher added value according to SAS® EVAAS® data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (1) Scores with asterisks (*) do 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 are still reported to both the teachers and their supervisors as they are here.

Teacher C's overall SAS® EVAAS® scores across years and subjects evidence that Teacher C detracted value from her students' learning (relative to all other HISD teachers) 100% of the time across three subject areas. This was likely because Teacher C taught some of the highest needs students, possibly across the district, however. 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 had been retained in grade one to four times prior.

Analyzing Teacher C's SAS® EVAAS® scores alongside her math PDAS scores was not possible as only two SAS® EVAAS® scores were available, although her social studies SAS® EVAAS® and PDAS scores were mildly related (r = 0.26). Teacher C's monetary bonuses and PDAS scores were also mildly related (r = 0.29). Until 2010-11, she "exceeded expectations" across almost every domain in terms of her supervisor evaluations. She was also given a "Teacher of the Year" award during the 2007-08 school year by her teacher peers.

Otherwise, the knowledge that Teacher C had about the SAS® EVAAS® was also limited. She understood that she was being compared to other HISD teachers who taught the same subject areas to students who were "very different than her students." She, like the others, never received training to understand, or professional development to improve, her value-added scores.

Teacher D

Teacher D graduated with a bachelor's degree in business and administration in 2005 and in 2007 was certified as a teacher for grades 4-8 via HISD's Alternative Teaching Certificate (ATC) program. She took a full-time teaching position in HISD in 2006. Illustrated in Table 4 is a summary of Teacher D's SAS® EVAAS® and PDAS scores and ASPIRE bonuses since 2007.

Table 4
Teacher D's SAS® 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

* Notes: Scores shaded as green indicate that the teacher added value according to SAS® EVAAS® data and in comparison to other similar teachers across the district. Scores shaded as red indicate the opposite. (1) Scores with asterisks (*) do 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 detectively different from the reference gain scores of other teachers across HISD given one standard error; however, the scores are still reported to both the teachers and their supervisors as they are here.

Up until 2009-2010 Teacher D, like Teacher A, switched back and forth across subject areas, demonstrating added overall from 2006-2009 50% of the time (3/6 SAS® EVAAS® observations) and demonstrating negative value 50% of the time (3/6 SAS® EVAAS® observations). According to her SAS® EVAAS® output, like Teacher A, 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 SAS® EVAAS® data (2009-2010), however, she seemingly detracted from student learning across all three subject areas. In 2009-2010 Teacher D was assigned to teach an inordinate number of ELLs who were transitioned into her classroom. This will be discussed in more detail later. Regardless, whether Teacher D demonstrated "a significant lack of student progress attributable to the educator," or "insufficient student academic growth reflected by [SAS® EVAAS®] value-added scores" is still disputable.

In terms of the relationship between Teacher D's performance on the PDAS and her students' SAS® EVAAS® scores in reading, there was a mild correlation (r = 0.29). In terms of her performance on the PDAS and her students' SAS® EVAAS® scores in language arts, there was a strong correlation (r = 0.92). In addition, the better Teacher D scored on the SAS® EVAAS® the more money she received (r = 0.79). Until 2010-11, she "exceeded expectations" or was "proficient" across every domain in terms of her supervisor evaluations.

In terms of Teacher D's knowledge about the SAS® EVAAS®, she reported not understanding how "they" could use different tests to evaluate her and whether she added or detracted value from her students' learning. She also did not trust whether "they" could really account for the types of students she had her in classroom, especially when she taught a disproportionate number of ELLs, in comparison and in her last year. While she reported having tried to figure SAS® EVAAS® out on her own online via the district's online resources, she found it very confusing. It "just did not hit home."

Findings

Reliability

According its developers, SAS® EVAAS® is meant to "assess and predict student performance with precision and reliability" and it is "the most robust and reliable" value-added system available, more than the "other simplistic models found in the market today" (SAS, 2011). In terms of the data presented here, however, it is clear that inconsistencies were a consistent problem. Across the four cases, issues with reliability were evident. Such issues with reliability are also well documented in the literature (Au, 2010; Baeder, 2010; Baker et al, 2010; CCSSO, 2010; Haertel, 2011; Koedel & Betts, 2007; Papay, 2010, Shaw & Bovaird, 2011; Schochet & Chiang, 2010).

Yet these four teachers were removed from their teaching positions "at least in part" due to SAS® EVAAS® data that in three of the four cases researchers evidenced as unreliable (see 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. In addition, as researchers suggest that at least three years of value-added data are needed to make such judgments (Brophy, 1973; Cody et al., 2010; Harris, 2011), and even then with a 25% risk of misclassification (Au, 2010; CCSSO, 2010; Otterman, 2010; Schochet & Chiang, 2010; Shaw & Bovaird, 2011), this is also troublesome. Not one of the four teachers had three years of consistent data (that were detectibly different from other similar teachers) to warrant non-renewal.

Other HISD teachers whom researchers interviewed (Collins, in progress) noted concerns about this as well, again comparing the receipt of merit monies based on SAS® EVAAS® data to "winning the lottery." 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." A social studies teacher added:

> 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 [SAS®] EVAAS® growth. My principal flipped him with the 6th grade science teacher who was getting the highest [SAS®] EVAAS® scores on campus. Huge [SAS®] 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.

SAS® EVAAS® developers claim they have evidence that teachers who move from one environment to another, even if radically different, continue to do as well and are classified the same in SAS® EVAAS® terms over time (LeClaire, 2011). Evidence presented herein should yield caution regarding this assertion.

Bias

Teachers credited such "chaos" to the different students they taught and the different classroom contexts in which they taught year-to-year. For example, while Teacher C's SAS® EVAAS® data illustrated that Teacher C consistently detracted value from her students' learning, and did so across subject areas, this was likely because Teacher C taught some of the highest-need students, possibly across the district.

In addition, HISD teachers note that those teaching inordinate numbers of special education students in mainstreamed classrooms are least likely to add value (Collins, in progress). Teachers teaching the same students over consecutive years (e.g., looping) report receiving bonuses for the first year and nothing the next as they are "maxing out" on growth, and actually "competing with themselves." Teachers agree that it is best for them "to get average kids, yes, because the regular kids, you can grow those kids!"

Teachers teaching gifted students report finding it very difficult to add value and get merit pay as a result (Collins, in progress; see also Wright, Horn, & Sanders, 1997). They report 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%.

Another 5th grade teacher working with gifted students explained:

> I have students [in a 5th grade gifted reading class] who score at the 6th, 7th, & 8th grade levels in reading. But I'm like please babies, score at the 9th grade level, cause if you don't score at the 9th or 10th grade or higher in 5th grade with me, I'm going to show negative growth. Even though you, you're gifted, and you're talented, and you're high! I can only push you so much higher when you are already so high. I'm scared.

Teachers teaching in grades in which ELLs were transitioned into mainstreamed English-only classrooms also report being the least likely to add value. 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 4th/5th grade teacher explained, "When they say nobody wants to do 4th grade – nobody wants to do 4th grade! Nobody" (Collins, in progress). This was evidenced in the data collected for Teacher D as well who, like Teacher A, switched back and forth across subject areas until her last year during which she purportedly detracted value across subject areas. This was the

year her supervisor assigned her to teach an ELL transition year, during which an inordinate number of ELLs entered her classroom.

Until SAS® EVAAS® developers can evidence that teachers teaching inordinate numbers of ELLs particularly in transition years, and teachers teaching special education or gifted students are not disparately impacted by the non-random placement of these students into their classrooms (Monk, 1987; Rothstein, 2009), terminating teachers on these grounds is remiss and morally indefensible. Just recently, both SAS® EVAAS® and HISD's Chief Human Resources Officer acknowledged via email that ceiling effects adversely impacted some teachers working with gifted students in their capacities to demonstrate value-added (A. Best, personal communication, January 21, 2012).

On the contrary, SAS® EVAAS® developers continue to claim that student background factors do not impact students' ability to grow year-to-year in the SAS® EVAAS® model, mainly because the system uses students' previous years of data as "blocking factors" to prevent such variables from biasing or distorting growth (Sanders & Horn, 1994, 1998; Sanders et al., 2009; Wright, White, Sanders, & Rivers, 2010). As evidenced here, these claims might not be entirely true. Appropriately, this is also one of SAS® EVAAS® developer's most highly contested claims (Amrein-Beardsley, 2008; Ballou, Sanders, & Wright, 2004; Braun, 2005; Cody et al., 2010; Kuppermintz, 2003; McCaffrey et al., 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004b; Sanders & Wright, 2008; Sanders et al., 2009; Tekwe, Carter, Ma, Algina, Lucas, Roth et al., 2004).

Teacher Attribution

The aforementioned lack of reliability could also be due to other context-related issues that further complicate the calculation of a teacher's value-added. Teacher B, for example, whose SAS® EVAAS® scores were negative for two years and positive for the most recent year for which she had data, taught for the same years alongside a math enrichment teacher who taught almost half of her students at the same time and an equal amount of time per week. Teacher A was not a teacher of record for approximately 50% of her students one of the years for which she was held accountable using the SAS® EVAAS® because she was moved from teaching the third to the fourth grade mid-year. Another HISD teacher taught alongside a reading specialist four days per week, and then posted the most growth and received the largest bonus she ever had (Collins, in progress). It is uncertain whether the reading specialist received a bonus for her apparent contributions as well.

Nonetheless, these instances raise concerns about the percentage of value teachers under this system add to, or detract from their students' learning and achievement and whether they can be held responsible for 100% of their students' scores. These issues might also play into why such inconsistencies are evident. Determining what percent of value-added scores can be attributed to teachers is very difficult, if even possible (Campbell & Stanley, 1963; Corcoran, 2010; Ishii, & Rivkin, 2009; Kane & Staiger, 2008; Kennedy, 2010; Linn, 2008; Nelson, 2011; Papay, 2010; Rothstein, 2009).

SAS® EVAAS® developers claim, though, that through a linking verification process (during which teachers mark for what percent of each student's instruction (s)he should be held accountable) they can partition out different teachers' value-added effects (Derringer, 2010; Sanders & Horn, 1994). However, there is no empirical evidence suggesting that numerically splitting or dividing teacher effects accurately accounts for a teacher's contribution. In addition, not only is such a practice counterintuitive, but breaking up effort across teachers using percentages and proportions is nonsensical given the interaction effects that occur among and between students and teachers (Monk, 1987). Teachers are situated in complex and collaborative learning environments. It is highly

unlikely their value-added effects can be fractionalized using simple or even complex mathematics and statistics.

Criterion-Related Evidence of Validity

One way to generate criterion-related evidence of validity is to assess whether teachers who demonstrate added value are also the teachers deemed effective through other, independent measures of teacher quality concurrently or at the same time (see also McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004a). In this instance, researchers examined whether the four non-renewed teachers also seemed to be ineffectual given their PDAS scores, specifically to determine if these teachers' supervisors also observed that these teachers were inadequate.

Analyzing the four teachers' SAS® EVAAS® scores over time alongside their PDAS scores, researchers found statistical signals indicating that both of these measures were not measuring the teaching effectiveness construct accurately and consistently across teachers. The better Teacher A did on the SAS® EVAAS® the worse she did in the eyes of her supervisor(s) ($r = -0.51$, $r = -0.83$, $r = -0.11$). Yet for Teacher B, the better she did on the SAS® EVAAS® the better she did on the PDAS ($r = 0.91$). This yields the type of correlation coefficient we would expect to see if in fact both indicators pointed in the same direction, yielding valid results. Researchers were not able to analyze Teacher C's math SAS® EVAAS® scores alongside her PDAS scores, but her social studies SAS® EVAAS® and PDAS scores were mildly related ($r = 0.26$). For Teacher D there were weak to strong results ($r = 0.29$, $r = 0.92$). The conclusion here is that there is nothing substantive to evidence that a valid teacher evaluation system, based on SAS® EVAAS® and PDAS scores, is in place and in use. This assertion is however limited by the small sample size herein.

Additionally, analyzing all four teachers' SAS® EVAAS® scores over time alongside their bonuses, researchers found that both of these measures failed to assess the teaching effectiveness construct accurately and consistently across teachers as well. The worse Teacher A did on the SAS® EVAAS® the more money she received ($r = -0.42$), and the better Teacher B did the more money she received ($r = 0.93$). Teacher C's monetary bonuses and SAS® EVAAS® scores were mildly related ($r = 0.29$), and Teacher D's monetary bonuses and SAS® EVAAS® scores were more strongly related ($r = 0.79$). Again, the small sample size certainly limits generalizability here, although evidence of this occurring elsewhere exists (Collins, in progress; Harris, 2011).

In addition, three of four teachers were honored with teaching awards (e.g., teacher of the year or month awards) during the same years for which they posted SAS® EVAAS® data that at least in part led to their contracts not being renewed. Teacher C ironically received a Teacher of the Year award, awarded to her by her peers, at the same time she detracted the most value from her students' learning according to her SAS® EVAAS® data. This raises additional concerns about whether these indicators are capturing the teaching effectiveness construct effectively, or validly. See notes above about sample size limitations.

External Pressures

It is also important to note that these teachers felt that they were targeted for termination because of the performance of the schools in which they taught, which were labeled "in-need-of-improvement" under NCLB. According to the four teachers, administrators were under intense district and state pressure, and administrators 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." Because they believed their supervisors perceived them to have low, or possibly lower value-added scores than their peers, the teachers felt that they had been put "on a list." It was at this time when they became most vulnerable, and when their PDAS observational scores plummeted.

Teacher A, for example, "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 she was 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.

Other teachers noted that their supervisors were beginning to skew their observational scores given external pressures to do so (Collins, in progress). 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 [SAS®] EVAAS®. So it looks like they're being consistent.

A middle school teacher agreed: "Well my evaluations were fine, but of course now they have to make the evaluation match the SAS® EVAAS®. We now have to go through that." An 8th grade teacher added:

> 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 that reflects the [SAS® 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. If you don't make [SAS® EVAAS®], then it goes against you in your PDAS. In a roundabout way she finds a way to put that against you." An 8th grade advanced English teacher added:

> My boss had to go to the district superintendent and explain why we needed to be kept, when ultimately the data showed that we weren't good teachers… [But] you've got other good teachers who are being thrown under the bus because of this system.

From these teachers' perspectives, it seems that many district administrators are more trusting of SAS® EVAAS® and are skewing PDAS data to match. This makes sense in theory, as the SAS® EVAAS® is the objective system that the district has purchased, and traditional observational scores are increasingly being dismissed as subjective (Harris, 2011). In Tennessee and New York there is evidence of local policies pushing such practices (Baker, 2012; Garland, 2012; Ravitch, 2012), although again such practices contradict the field standards encouraging the use of multiple measures for decision-making (AERA, APA, & NCME, 2000).

Diagnostics and Formative Uses

Overall, Teachers A, B, C, and D were only familiar with SAS® EVAAS®. They understood that they were being compared to other similar teachers within the district, and they understood their scores were available each year via the district's online portal system, but that was about the extent of their knowledge. Nobody had explained their SAS® EVAAS® data to them, and none of

the four teachers understood what their SAS® EVAAS® numbers meant, how they were calculated, how their SAS® EVAAS® scores could be "used against [them]," or conversely how they could use their SAS® EVAAS® scores to help them improve their instruction. Teacher D took steps to figure out her SAS® EVAAS® scores on her own, but her SAS® EVAAS® scores still "just did not hit home" (see also Eckert & Dabrowski, 2010).

The four terminated teachers did not receive professional development from HISD or SAS® EVAAS® as a result of their value-added scores either. However, given the scores illustrated in Tables 1-4, whether each teacher needed professional development to improve their value-added scores is disputable. Because they were terminated at least in part due to their SAS® EVAAS® scores, and because they were reportedly not given professional development to improve their scores, this too is troublesome. Similarly, none of the teachers noted that they used SAS® EVAAS® data to inform their instruction, in many ways because they did not understand it.

In short, no data suggest that for these four teachers in HISD that the SAS® EVAAS® system "provides valuable diagnostic information about [instructional] practices," helps educators become more proactive and make more "sound instructional choices," and helps teachers use their "resources more strategically to ensure that every student has the chance to succeed" (SAS, 2011). In addition, 60% of a sample of HISD teachers indicate that they are not using SAS® EVAAS® data to inform their instruction either. This is not to say, however, that this is not occurring elsewhere, perhaps in the district for the other 40% (taking into account sampling error) or in other states, districts, and schools using the SAS® EVAAS® system (see, for example, marketing testimonials available on the SAS website, SAS, 2012).

Conclusions

In the end, Teachers A, B, and D pursued due process hearings, but they decided not to follow their hearings through to culmination. They ultimately decided to quit teaching in HISD or altogether. Teacher C (the teacher who according to her SAS® EVAAS® output had the poorest 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.

But in sum, and based on the cases of these four teachers, it seems the district is inappropriately using inconsistent data within and across subject areas to make high stakes decisions about teachers, and in this case teacher termination. This was evidenced through examinations of four teachers' SAS® EVAAS® data, how they correlated with other data meant to capture the same teaching effectiveness construct, and teachers' complementary stories, collected to better examine the data and other relevant issues.

The goal of this study was also to examine other intended and unintended effects of the SAS® EVAAS® system, in particular given HISD's use of the system for high-stakes decision-making. In terms of intended effects, the four terminated teachers did not seem to understand SAS® EVAAS® output well enough to understand or use value-added data to inform or improve their instruction. This happens particularly when district leaders do not provide professional development to promote formative use (see also Eckert & Dabrowski, 2010; Harris, 2011). But this is also particularly problematic in that "when cases challenging dismissal based on VAM make it to court, deliberations will center on [among other things]…whether teachers are able to understand the basis for which they have been dismissed and whether it is assumed that they

have had any control over their fate" (Baker 2012). In general, whether VAMs succeed in their intended objectives will also bet contested. Researchers examined these issues here by framing this study around the marketing materials publicized by SAS® EVAAS®.

In terms of unintended effects, however, researchers also evidenced specific issues with reliability, bias, teacher attribution, and validity; issues also evident in the growing research literature and also named in the anticipated lawsuits (Baker, 2012). Researchers found that high-stakes use of SAS® EVAAS® in this district seems to be exacerbating unintended effects.

Results from the four teachers indicate there are consistent problems with inconsistencies with the SAS® EVAAS® data (see also Au, 2010; Baeder, 2010; Baker et al., 2010; Corcoran, 2010; Haertel, 2011; Koedel & Betts, 2007; Papay, 2010). These inconsistencies are likely related to the measurement errors already inherent in standardized tests and the errors intensified when SAS® EVAAS® researchers mix norm- and criterion-referenced tests together, use tests that are not appropriately scaled or designed to measure growth upwards, and try to account for or inpute missing longitudinal data. SAS® EVAAS® researchers also do not seem to be sufficiently controlling for many extraneous variables using even their most sophisticated controls and blocking methods. Such extraneous variables include parental contributions to learning outside of school, after school programming, pullout and intensive programs, tutor effects, prior teachers' residual effects on current year test scores, differential summer learning losses and gains, student motivation factors, peer and teacher interaction effects, and other variables impacting non-traditional classrooms (see also Haertel, 2011; Harris, 2011; Rothstein, 2009; Sanders et al., 2009; Shaw & Bovaird, 2011; Wilson, Hallman, Pecheone, & Moss, 2007).

These inconsistencies are also likely related to the proposition that SAS® EVAAS® output are biased by student demographics. This was evidenced in this study, particularly for the teachers who taught ELLs and an inordinate number of students previously retained in grade (see also Newton et al., 2010; Hill et al., 2011; Rothstein, 2009). HISD teachers also mentioned not wanting to teach high numbers of gifted, special education, or ELL students for fear of posting low SAS® EVAAS® scores (Collins, in progress). In addition, the issue of SAS® EVAAS® bias seems to hold true for teachers of high achieving or gifted students when ceiling effects prevent their students' aggregated scores from yielding significant growth (see also Wright, Horn, & Sanders, 1997). SAS® EVAAS® methodologists have recently verified that test ceilings are a concern as well, without yet providing suggestions about how to address this issue. Inversely, researchers have no evidence to date that regression to the mean artificially inflates value-added scores for teachers with large groups of low-scoring students.

Limited evidence also exists to indicate that SAS® EVAAS® output are related to at least one other correlated criterion (i.e., evidence of criterion-related validity), in this case in terms of the PDAS (see also Milanowski et al., 2004; Wilson, Hallman, Pecheone, & Moss, 2007). It is methodologically and pedagogically more beneficial that a teacher be classified similarly on at least one other, medium-to-highly correlated, unbiased measure to independently assess the same construct at the same time before consequences are tied to value-added output (AERA, 2000). And this must happen before anyone can make a solid case that a teacher is effective or ineffective, or should be monetarily rewarded or contractually terminated (Baker et al., 2010; Harris, 2011; Hill, 2009; Hill et al., 2011; Newton et al., 2010; Papay, 2010). The more that multiple indicators converge or correlate (e.g., in terms of inter-indicator consistency; see, for example, Amrein-Beardsley, Haladyna, & Polasky, 2012), and the more years over which the indicators yield the same results, the stronger the accountability system should be, and the more justifiable high-stakes decision(s) surrounding teacher evaluation should become.

Either way, high-stakes decisions should not be made on the basis of value-added scores alone (AERA, APA, & NCME, 2000). The evidence presented here indicates that, at least in the cases of these four teachers, HISD is violating this highly relevant standard calling for multiple indicators, or distorting it as principals seem to be skewing at least some teachers' PDAS scores to match what appear to be the superior scores derived via SAS® EVAAS® (see also Baker, 2012; Garland, 2012; Ravitch, 2012).

Whether those at SAS® EVAAS® should share in the responsibility to ensure their system is used properly is a debate for another day. Perhaps the focus of such conversations need to shift towards discussing how such system are to be used, and whom should be held responsible for ensuring they are used correctly and validly. While in this case it would be easy to blame the for-profit institution netting significant returns from model sales, perhaps it is not SAS's responsibility to ensure proper use of the SAS® EVAAS®. However, SAS® EVAAS® does have the responsibility of identifying effective teachers, schools, and systems, in a precise, unbiased, and reliable manner and providing "valuable diagnostic information about [instructional] practices," helping educators become more proactive and make more "sound instructional choices," and helping teachers use "resources more strategically to ensure that every student has the chance to succeed" (SAS, 2012). These deliverables are advertised in the SAS® EVAAS® literature and marketing materials. Yet these claims were countered with empirical, albeit case-based evidence in this study. Researchers further situated these findings in the ever-evolving literature base surrounding VAMs, as well as experiences from other HISD teachers (Collins, in progress).

In theory, VAMs allow for richer analyses of test score data because groups of students are *simply* followed to assess their learning trajectories from the time they enter a classroom to the time they leave. In practice, however, these models do not seem to work in the ways purported, and in this case, advertised. This was evidenced here, as researchers conducted one of the first studies to examine how this particular value-added system, as marketed, is working in practice. This is also the first study to look at this particular value-added system and its implications from the teacher's perspective.

We ultimately assert that even though results may not generalize far beyond the confines of this study, there is a lot to be learned, given the results presented, about the real impact of the SAS® EVAAS® on the very real lives of some teachers in Houston. Perhaps the methodologists pushing, and in this case selling the SAS® EVAAS® model for profit, are promising more than their model can and ever will deliver. What they are delivering, however, is also a series of unintended consequences, some of which are being exacerbated in HISD with its highly consequential use of SAS® EVAAS® output. These unintended consequences cannot continue to go unrecognized, and whether the unintended consequences outweigh the intended consequences warrants further research and evaluation.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2000). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association (AERA) (2000). *Position statement on high-stakes testing in PreK-12 education.* Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher, 37*(2), 65-75. doi: 10.3102/0013189X08316420

Amrein-Beardsley, A. Haladyna, T, & Polasky, S. A. (2012). *Imagining a new system: Using multiple measures and inter-indicator consistency to measure value-added.* Paper presented at the annual convention of the American Association of Colleges for Teacher Education (AACTE), Chicago, IL.

Au, W. (2010, Winter). Neither fair nor accurate: Research-based reasons why high-stakes tests should not be used to evaluate teachers. *Rethinking Schools*. Retrieved from http://www.rethinkingschools.org/archive/25_02/25_02_au.shtml

Baeder, J. (2010, December 21). Gates' measures of effective teaching study: More value-added madness. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/on_performance/2010/12/gates_measures_of_effective_teaching_study_more_value-added_madness.html

Baker, B. D. (2012, March 31). Firing teachers based on bad (VAM) versus wrong (SGP) measures of effectiveness: Legal note. *School Finance 101*. Retrieved from http://schoolfinance101.wordpress.com/2012/03/31/firing-teachers-based-on-bad-vam-versus-wrong-sgp-measures-of-effectiveness-legal-note/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+GLCWorthARead+%28Worth+A+Read%29&blogsub=confirming#subscribe-blog

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

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66. doi:10.3102/10769986029001037

Banchero, S. & Kesmodel, D. (2011, September 13). Teachers are put to the test. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424053111903895904576544523666669018.html?mod=googlenews_wsj

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10(*3), 245–252. doi:10.2307/1161888

Campbell, D. (1975). Degrees of freedom and the case study. *Comparative Political Studies, 8*, 178-185.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.

CCSSO. (2010, October). *Teacher evaluation literature review*. Washington, DC: Council of Chief State School Officers.

Cody, C. A., McFarland, J., Moore, J. E., & Preston, J. (2010, August). *The evolution of growth models*. Public Schools of North Carolina. Retrieved from http://www.dpi.state.nc.us/docs/intern-research/reports/growth.pdf

Collins, C. (in progress). Houston, we have a problem: Studying the SAS Education Value-Added Assessment System (SAS® EVAAS®) from teachers' perspectives in the Houston

Independent School District. (Unpublished doctoral dissertation). Arizona State University, Tempe.

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

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating. Quantitative and qualitative research.* (3rd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Day, C., Sammons, P., & Gu, Q. (2008). Combining qualitative and quantitative methodologies in research on teachers' lives, work, and effectiveness: From integration to synergy. *Educational Researcher, 37*(6), 330-342. doi:10.3102/0013189X08324091

Derringer, P. (2010, August). RTT in Tennessee: Assessment done right. *Technology and Learning, 31*(1), 40. Retrieved from http://www.techlearning.com/article/31572

Eckert, J. M., & Dabrowski, J. (2010, May). Should value-added measures be used for performance pay? *Phi Delta Kappan, 91*(8), 88-92.

Feagin, J., & Orum, A. (1991). *A case for the case study*. Chapel Hill, NC: The University of North Carolina Press.

Flyvbjerg, B. (2011). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219-245. doi:10.1177/1077800405284363

Garland, S. (2012, February 7). Tennessee teacher evaluation systems have rough road ahead. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/2012/02/07/tennessee-teacher-evaluat_n_1260790.html?page=1

Gerring, J. (2004). What is a case study and what is it good for? *The American Political Science Review, 98*(2), 341-354. doi:10.1017/S0003055404001182

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Greene, J. C., Caracelli, V. J., & Graham, W. D. (1989). Toward a conceptual framework for mixedmethod evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco, CA: Jossey-Bass.

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.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 693-700. doi:10.1002/pam.20464

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Hill, H. C. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 700-709.

Hill, H. C., Kapitula, L., & Umland, K. (2011, June). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi:10.3102/0002831210387916

Ho, A. D., Lewis, D. M., & Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15-26. doi:10.1111/j.1745-3992.2009.00159.x

Houston Independent School District (HISD). (2010). *Accelerating Student Progress and Increasing Results and Expectations (ASPIRE) at HISD: A guide for parents and our community*. Retrieved from www.houstonisd.org/.../PDF/ASPIRE_ParentGuide_2010114web.pdf

Houston Independent School District (HISD). (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*(4), 520-536. doi:10.1162/edfp.2009.4.4.520

Jacob, B. A., & Lefgren, L. (2007, June). *Can principals identify effective teachers? Evidence on subjective performance evaluation in education*. Retrieved from econ.byu.edu/faculty/Lefgren/Assets/papers/principals.pdf

Johnson, R. B. & Onwuegbuzie, A. J. (2004) Mixed methods research: A research paradigm whose time has come. Educational Researcher, 33(7), 14-26. doi:10.3102/0013189X033007014

Kane, T. J., & Staiger, D. O. (2008, March). *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.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. doi:10.3102/0013189X10390804

Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function*. Working Paper No. 2007-03. Nashville, TN: National Center on Performance Initiatives.

LeClaire, B. (2011, June 1). 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.

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40,* 699-711.

Lowrey, A. (2012, January 6). Big study links good teachers to lasting gain. *The New York Times*. Retrieved from http://www.nytimes.com/2012/01/06/education/big-study-links-good-teachers-to-lasting-gain.html?emc=eta1#comments

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003*). Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004a). Let's see more empirical studies on value-added modeling of teacher effects: A reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics, 29*(1), 139–143.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101. doi:10.3102/10769986029001067

Mellon, E. (2010, January 14). 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/disp/story.mpl/metropolitan/6816752.html

Milanowski, A., 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 Series TC-04-01, Madison, WI: University of Wisconsin-Madison, Wisconsin. Center for Education Research, Consortium for Policy Research in Education.

Monk, David H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal, 88*(2), 167–187. doi:10.1086/461531

Nelson, F. H. (2011, April). *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., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23). Retrieved from http://epaa.asu.edu/ojs/article/view/810

Otterman, S. (2010, December 26). 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*(1), 163-193. doi:10.3102/0002831210362589

Patton, M. Q. (2001). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

PDAS. (2004). *Professional Development Appraisal System (PDAS).* Austin, TX. Education Service Center Region XIII. Retrieved from http://www5.esc13.net/pdas/

Race to the Top. (2009, November). *Race to the Top program: Executive summary*. Washington, DC: US Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Ragin, C. C. & Becker, H. S. (2000). Cases of "what is a case?" In C. C. Ragin & H. S Becker, *What is a case? Exploring the foundations of social inquiry,* pp. 1-17. Cambridge, UK: The Press Syndicate of The University of Cambridge.

Ravitch, D. (2012, February 21). No student left untested. *The New York Review of Books.* http://www.nybooks.com/blogs/nyrblog/2012/feb/21/no-student-left-untested/

Robelen, E. W. (2012, January 9). Yardsticks vary by nation in calling education to account. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2012/01/12/16testing.h31.html?tkn=ZRXFgiQ5krPVo%2FsHmf1v%2Bh33GqSq%2ByE1LBEQ&cmp=ENL-EU-NEWS1&intc=EW-QC12-ENL

Rothstein, J. (2009, January 11). *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

Sanders, W. L. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329-339.

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*(3), 247-256.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Accountability System: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L., & Wright, S. P. (2008, April 14). *A response to Amrein-Beardsley (2008): "Methodological concerns about the Education Value-Added Assessment System."* Available: 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, November). *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. (2007). *Resource guide for value-added reporting.* Cary, NC. SAS Institute. Retrieved from www.dpi.state.nc.us/docs/evaas/guide/resourceguide.pdf

SAS. (2012). *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

Schochet, P. Z. & Chiang, H. S. (2010, July). *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/

Shaw, L. H. & Bovaird, J. A. (2011, April). *The impact of latent variable outcomes on value-added models of intervention efficacy.* Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Sparks, S. D. (2011, November 15). Value-added formulas strain collaboration. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/11/16/12collab-changes.h31.html?tkn=OVMFb8PQXxQi4wN6vpelNIr7%2BNhOFCbi71mI&intc=es

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36. doi:10.3102/10769986029001011

Thomas, G. (2011a). A typology for the case study in social science following a review of definition, discourse, and structure. *Qualitative Inquiry, 17*(6), 511-521. doi:10.1177/1077800411409884

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

Yin, R. (1994). *Case study research: Design and methods* (2nd ed.). Beverly Hills, CA: Sage Publishing.

About the Authors

Audrey Amrein-Beardsley, Ph.D.
Mary Lou Fulton Teachers College
Arizona State University
Email: audrey.beardsley@asu.edu

Dr. Amrein-Beardsley is currently an Associate Professor in the Mary Lou Fulton Teachers College at Arizona State University. Audrey's research interests include educational policy, research methods, and more specifically, high-stakes tests and value-added measurements and systems. In addition, she researches aspects of teacher quality and teacher education. She is also the creator and host of a show titled Inside the Academy during which she interviews some of the top educational researchers in the academy. For more information please see: http://insidetheacademy.asu.edu.

Clarin Collins, Ph.D. Candidate
Mary Lou Fulton Teachers College
Arizona State University
Email: clarin.collins@asu.edu

Clarin Collins is a doctoral candidate in Educational Leadership and Policy Studies at Arizona State University. Her research interests include national and local policy implementation at the classroom level, teacher influences on policy making and implementation, and education evaluation and accountability systems. Her dissertation study analyzes teachers' understanding of and experiences with the SAS Educational Value-Added Assessment System (SAS® EVAAS®) in a large, urban, high-needs school district using the SAS® EVAAS® to evaluate teachers with high-stakes consequences.

education policy analysis archives

Volume 20 Number 12 April 30th, 2012 ISSN 1068-2341