SPECIAL ISSUE
Value-Added: What America's Policymakers Need to Know and Understand

education policy analysis archives

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

epaa aape

Arizona State University

Volume 21 Number 7 January 29, 2013 ISSN 1068-2341

Value-added Teacher Estimates as Part of Teacher Evaluations: Exploring the Effects of Data and Model Specifications on the Stability of Teacher Value-added Scores

Nicole B. Kersting
Mei-kuang Chen
University of Arizona
United States of America

James W. Stigler
University of California, Los Angeles
United States of America

Citation: Kersting, N. B., Chen, M., & Stigler, J. W. (2012). *Education Policy Analysis Archives, 21*(7). This article is part of EPAA/AAPE's Special Issue on *Value-Added: What America's Policymakers Need to Know and Understand*, Guest Edited by Dr. Audrey Amrein-Beardsley and Assistant Editors Dr. Clarin Collins, Dr. Sarah Polasky, and Ed Sloat.
http://dx.doi.org/10.14507/epaa.v21n7.2013

Abstract: If teacher value-added estimates (VAEs) are to be used as indicators of individual teacher performance in teacher evaluation and accountability systems, it is important to understand how much VAEs are affected by the data and model specifications used to estimate them. In this study we explored the effects of three conditions on the stability of VAEs and evaluated their relative impact. We varied how we accounted for differences among students in their prior learning; whether we

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 7/6/2012
Revisions received: 9/6/2012
Accepted: 12/31/2012

estimated VAEs from single or multiple cohorts of students; and the number of students contributing to the VAE for each teacher. Using data from one of the largest school districts in the nation, we created a single, complete data set and used it to estimate several sets of VAEs for each of the 3651 5th grade mathematics teachers. We found that approximately two thirds of teachers' were stable in that they remained in the same performance group across all conditions. We also found that differences in number of students used for VAEs accounted for up to one third of teacher reclassifications into different performance groups; single versus multiple cohort models accounted for about one fifth; and different methods for controlling for student prior learning accounted for about one sixth (16%) of teacher reclassifications. We relate our findings to characteristics of our data and discuss implications for educational policy.
Keywords: Value-added Analysis, Value-added Models; Value-added Estimates; Stability; Teacher Evaluations; Accountability.

Las estimaciones de valor agregado docente como parte de las evaluaciones docentes: Explorando los efectos de los datos y las especificaciones de los modelos en la estabilidad de las puntuaciones de valor agregado docentes

Resumen: Si estimaciones de valor añadido docente (VADs) van a ser utilizadas como indicadores de rendimiento de los docentes en cada evaluación de los profesores y en los sistemas rendición de cuentas, es importante entender cómo los VADs son afectados por los datos recogidos y las especificaciones de los modelos utilizados para estimarlos. En este estudio hemos explorado los efectos de tres condiciones para la estabilidad de VADs y evaluar sus impactos relativos. Hemos variado cómo se estimaron las diferencias entre los estudiantes en sus aprendizajes previos; hemos estimado VADs de una o múltiples promociones de estudiantes; y el número de estudiantes que contribuyen a la de cada uno de los VADs de los profesores. Utilizando los datos de uno de los mayores distritos escolares de la nación, hemos creado un conjunto único de datos completo y lo utilizamos para estimar varios conjuntos de único para cada uno de los 3651 profesores de matemáticas de 5to. Grado. Encontramos que aproximadamente dos terceras partes de los maestros se mantuvieron estables y que permanecieron en el mismo grupo de rendimiento a lo largo de todas las condiciones. También encontramos que las diferencias en el número de alumnos utilizados para VADs representaron hasta un tercio de las reclasificaciones de los docente en diferentes grupos de rendimiento; grupos individuales frente a múltiples representaron aproximadamente una quinta parte, y diferentes métodos de control para los aprendizajes previos de los estudiantes representaron cerca de un sexto (16%) de las reclasificaciones de los docentes. Relacionamos nuestros hallazgos a las características de nuestros datos y analizamos sus implicaciones para las políticas de educación.
Palabras clave: modelos de valor añadido (MVA); estimaciones de valor añadido; estabilidad; evaluaciones docentes; rendición de cuentas.

Estimativas do valor adicionado dos professores como parte das avaliações docentes: Explorando os efeitos dos dados e especificações dos modelos na estabilidade das pontuações de valor agregado docente.

Resumo: Se estimativas do valor agregado docente (VADs) são utilizados como indicadores de desempenho dos professores, em avaliações dos professores e nos sistemas de prestação de contas, é importante entender como as VADs são afetados pelos dados coletados e as especificações dos modelos utilizados para estimá-los. Neste

estudo, exploramos os efeitos de três condições para a estabilidade dos VADs para avaliar seus impactos relativos. Nós mudamos a forma como se estimaram as diferenças entre os alunos em suas aprendizagens anteriores; Se as VADs eram estimadas para coortes de estudantes únicas ou múltiplas; é o número de estudantes que contribue para cada um dos VADs dos professores. Usando dados de um dos maiores distritos escolares do país, criamos um conjunto único de dados completos para estimar VADs para cada um dos 3.651 professores de matemática de 5º ano. Descobrimos que cerca de dois terços dos professores eram estáveis e permaneceram no mesmo grupo de desempenho em todas as condições. Também descobrimos que as diferenças no número de alunos usados para VADs representaram até um terço de reclassificações em grupos diferentes de desempenho; grupos individuais versus múltiplos foram responsáveis por cerca de um quinto das reclassificações; e diferentes métodos de controle das aprendizagens prévias dos estudantes responderam por cerca de um sexto (16%) das reclassificações dos professores. Nós relacionamos nossos resultados com as características dos dados e analisamos as implicações para as políticas de educação.

Palavras-chave: modelos de valor agregado (MVA): estimativas de valor agregado; estabilidade; avaliações dos professores; prestação de contas.

Introduction

Value-added scores as a measure of individual teacher performance have garnered a lot of attention in recent years. Encouraged by educational policy decisions, an increasing number of states and districts have adopted the use of value-added measures as part of formal teacher evaluation and accountability systems, including performance-based pay (National Council on Teacher Quality, 2011; Sanders & Rivers, 1996; Sanders, Saxton, & Horn, 1997). Under new legislation, states can obtain waivers from key provisions specified in *No Child Left Behind* (NCLB, 2001), among them that all students be proficient in reading and mathematics by 2014, in exchange for developing measure-based evaluation and support systems to improve teacher effectiveness. Acceptable use of teacher value-added estimates (hereafter VAEs) in high-stakes decision contexts requires that their accuracy and meaning are well understood (Hill, Kapitula & Umland, 2011; Measures of Effective Teaching Project, 2010).

There is little question that teachers have an impact on student learning (Rowan, Miller, & Correnti, 2002; Hatti, 2003). Yet, quantifying that impact for evaluation or accountability purposes in ways that rise above the local context and ensure fairness for teachers has remained a challenge. Evidence from past efforts has been mixed. Principal ratings have often been shown to be overly lenient, providing little differentiation among teachers (The New Teacher Project, 2007), although other studies have found them to be fairly accurate at the two ends of the teacher effectiveness distribution (Milanoswski, 2004). Some accountability policies have focused on the school as the unit of improvement (NCLB, 2001).

Value-added models (hereafter VAMs), at least in theory, provide an opportunity for improvement over the status quo because they seek to estimate teachers' contributions to student learning for a given year, taking prior learning into account. Although currently VAMs rely primarily on standardized test scores to assess student learning, which might not accurately measure all learning that takes place in classrooms and estimated teacher performance might vary depending on the test (Papay, 2011), they do differentiate among teachers because each teacher's performance is estimated in comparison to the average

performance in a district or state. In addition, VAEs might provide standardization and transparency, if models can successfully isolate the actual "teacher effect" from all those sources affecting student learning that teachers often have little or no control over, such as differences in students prior learning and their backgrounds, classroom composition, or school functioning.

Although commonly referred to as teacher effects or teacher performance estimates, VAEs are by no means causal effects (Raudenbush, 2004; Rothstein, 2007); rather estimated differences between expected and observed student learning are attributed to the classroom teacher, assuming that other potential sources affecting student learning are either accounted for in the model or are sufficiently small. If teacher VAEs are to be used as an indicator of individual teacher performance in teacher evaluations and accountability systems, it is important for policy makers to ask how much VAEs are affected by the particular models and data used to estimate them. If VAEs vary substantially as a function of the models and data used to estimate them, it might be an indication that such estimates are fairly noisy and that they might not be suitable for use in high-stakes decision contexts. On the other hand, if VAEs are fairly stable across a wide range of modeling conditions, it might be an indication that they reflect a somewhat stable aspect of teacher performance.

Over the past decade, a growing number of VAMs have been developed and different model specifications have been explored. While those who develop and study these models have engaged in a spirited discussion about the relative benefits and limitations of different approaches and the overall quality of value-added scores, the technical nature of the discussion might have left those who need to evaluate and legislate their use, policy makers and district administrators, wondering just how much data and modeling differences matter. Are researchers and statisticians splitting hairs about technical details that result in little or no practical differences in value-added scores, or are such effects substantial, suggesting that using VAEs as a measure of teacher performance in high-stakes accountability systems might be problematic? If teacher performance depends to a large extent on how performance is estimated and if the same teacher might be evaluated differently under different models, it would have important implications for the use of such scores.

To address this larger question, we explored the effects of three basic conditions on VAEs, asking the following questions: (1) What are the effects of different statistical controls on the stability of teacher VAEs and performance group designations? (2) How stable are VAEs from year to year and over multiple years, and how do single cohort VAEs compare to VAEs estimated from multiple cohorts? (3) What is the effect of student sample size on the precision (i.e., the size of standard errors) of VAEs and on teachers' performance group designations? We focused on these three areas because they are central to value-added approaches. To answer each research question we investigated the relative impacts of the three conditions on changes in teachers' relative standing (as correlations) and on the percentage of teachers who either changed or remained in their respective performance groups. We differentiated teachers into three performance groups: significantly below average, average performance, and significantly above average (Ballou, Sanders, & Wright, 2004). This decision is somewhat arbitrary, but we reasoned that teachers who are estimated to perform different at statistically significant levels from the average as measured by value-added might differ in important other ways, for example, instructional profiles that might be related to their performance.

Being able to quantify the effects and relative impact of different model specifications and data conditions on teacher VAEs represents an important step for policy

makers to understand the underlying factors that might determine if and under which particular conditions (in a district or state) their use in accountability systems might or might not be reasonable. It is conceivable that in some districts or states, VAEs can be shown to be sufficiently stable for a large majority of teachers, so that they could be used as one of several measures of teacher performance, while that might not be the case in other places. To begin exploring the effects of data and model specification on VAEs we obtained data from one of the largest school district in the nation and created a single data set for our analyses that contained no missing data to ensure that any observed variability in VAEs was only due to changes in the three conditions under study.

Literature Review

Over the last decade, a growing number of VAMs have been developed, and different model specifications have been explored, with the dual goal of understanding and improving the quality of teacher value-added scores (Ballou, 2005; Ballou et al., 2004; Briggs & Domingue, 2011; Buddin, 2010; Harris & Sass, 2010; Harris, Sass, & Semykina, 2012; Harris, 2011; Hill, et al., 2011; Kane & Staiger, 2002a; 2002b; 2008; McCaffrey et al., 2004; McCaffrey, Sass, Lockwood, Milhaly, 2009; Papay, 2011; Rothstein, 2007; Schochet & Chiang, 2010). Yet, understanding which data and model specifications have a larger effect on teacher VAEs compared to others has been challenging because the number of model specifications and their possible combinations is too large to be fully investigated in any single study. Often, studies report results from only one or another particular value-added analysis (Buddin, 2010; Noell & Burns, 2006; Roderick, Jacob, & Bryk, 2002). Others focus on a single model specification (Ballou et al., 2004;). Yet others change several model specifications at once (Harris, Sass, & Semykina, 2012), for example, when reanalyzing data to call into question results obtained from a previous analysis (Buddin, 2010; Briggs & Domingue, 2011), which makes it difficult to understand the relative impact of any single specification change on VAEs. The number of studies that systematically explore the relative impact of different conditions on VAEs is limited. A recent study by Papay (2011) found that the effects that different outcome measures (tests) had on VAEs were larger than the effects school and other student background characteristics.

There is also no commonly shared metric to report on the stability of teacher VAEs, which makes comparing results across studies challenging. Studies that examine the impact of model specifications on teacher VAEs describe stability in terms of correlations in teachers' relative standing across time, or as changes in teacher classification into performance groups. Different studies divide the value-added distribution into different numbers of performance groups. Some studies specify three groups: average effective and significantly above or below average effective, (Ballou et al., 2004). Others use four (Papay, 2011) or five groups respectively (McCaffrey et al., 2009), based on quartiles or quintiles , to report the percentage of teachers whose group designation either remained the same or changed. As a consequence, comparing percentage changes in teacher designations are difficult to interpret when those percentage changes are based on different numbers of performance groups.

These kinds of issues can make it challenging for users of teacher value-added scores, such as districts and states, to understand the trade-offs between specific model specifications and their relative impact on teacher VAEs. Yet, understanding which data and model specifications might optimize the quality of VAEs in one district or state but not in another is of high practical importance. For example, knowing whether the number of

students per teacher has a greater effect on the stability of VAEs than the number of years for which prior test scores are available for each student is important information for building VAMs because the number of students typically decreases as the number of grades for which test scores are required increases. It might also be valuable to know if in large urban districts with high student mobility, such as the one in our study, controlling for student prior learning using a combination of previous year test scores and additional student background variables is comparable to controlling with student test scores from multiple prior grades, if the former increases the number of students that can be included in a teacher's value-added estimate. We focused on statistical controls, single versus multiple cohort models, and the number of students per teacher because they appeared central, basic, and inter-related. Below we briefly summarize findings on the stability of VAEs pertaining to each of the three conditions under study.

Accounting for Differences in Students, Classrooms, and Schools

Isolating the teacher effect, or teachers' contribution to student learning in VAMs, has remained a challenge. Teacher effects as causal effects require random assignment into treatment groups, in this case, of students to teachers, classrooms, and schools because student learning outcomes are likely to vary independently of the teacher students had in a given year (Raudenbush, 2004). In absence of random assignment the teacher effect in value-added approaches is based on attribution, assuming that differences between expected and observed student learning are primarily due to the teacher. If students exceeded their expected learning over a given year, their teacher is assumed to have added value. If students remained below expectations, the teacher is assumed to have not. Because in reality students are not randomly assigned to teachers, classrooms, and schools, a causal attribution requires that experimental conditions, afforded by random assignment, can be approximated (Raudenbush, 2004). This is commonly done by accounting for differences statistically, rather than by design. However, when statistically controlling for differences in students, classrooms and schools we not only assume that control variables such as student prior learning and background characteristics are unrelated to teacher assignment, although it could be that some teachers are particularly effective with some student populations but not others (McCaffrey et. al., 2004; Rothstein, 2007), but we are also limited by the measures that are available to adjust for differences and can be included in the model. Any unmeasured sources of systematic variation that can't be accounted for in the model might remain part of the between-teacher variance or the VAEs and hence, of the ascribed teacher effect. Because VAEs are essentially the left over unexplained between-teacher variance, how to control and how much to control in VAMs has been a central area of inquiry and research (Ballou et al., 2004; Hill et. al, 2011; McCaffrey et al., 2004).

Studies investigating statistical controls in VAMs have primarily focused on two issues: (1) How can VAMs account for student learning differences among students, classrooms, and schools sufficiently and efficiently so that VAEs contain little or no effects of sources of variation unrelated to the teacher? (2) What are the appropriate measures (i.e., variables) that should be used in VAMs to control for those differences that are unrelated to teacher performance?

One key concern in VAMs is that some students are likely to learn more than others due to their background, not because of the teacher who is teaching them. Unless the playing field is leveled by sufficiently controlling statistically for those differences some teachers might obtain VAEs that over- or under-estimate their contributions to student learning. With limited statistical controls or poor measures, teacher VAEs tend to be difficult

to interpret because teachers estimated contribution to learning is confounded with all those sources of variation known to affect student learning that could not be sufficiently controlled for. If for example, only students' previous year test scores are available to control for prior student learning and if test scores are known to be unreliable measures of student learning, VAMs might not appropriately account for difference in prior learning, which directly affects the VAEs. That is why some have recommended that student scores from three prior grades be included in VAMs (Ballou et al., 2004)), a condition which might seriously limit the number of students whose records can be used in VAMs and the grade levels for which teacher VAEs might be estimated. With adequate statistical controls, proponents of value-added approaches argue, most of the variance in student learning stemming from differences in students, classrooms, and schools that is unrelated to teacher performance can be reduced and teachers' contribution to student learning can be estimated more accurately (Sanders & Rivers, 1996; Sanders, Saxton, & Horn, 1997).

The flip side of not controlling sufficiently for differences in students is to control too much. Students might be systematically assigned to teachers, relating student characteristics meaningfully to teacher performance. For example, not all teachers might be equally effective teaching all kinds of students. Some teachers might be particularly effective with English language learners (ELLs), perhaps due to special training, while others might be more effective teaching gifted students. If sorting of students to teachers is strategic and deliberate, statistically controlling for student differences might create bias because teacher performance is estimated assuming that students across all teachers are statistically equivalent (McCaffrey, et al., 2004).

There are two basic approaches to control in VAMs: Models that rely exclusively on student test scores from multiple prior grades and prior teacher effects as inputs to estimate teacher contribution to current learning (e.g., the Education Value-added Assessment System [EVAAS]); and models that adjust for student, classroom, and school differences by including a combination of students' previous-year test scores and other background characteristics known to be related to student learning and their classroom and school level aggregates or other measures of classroom and school-level inputs if available (Buddin, 2010; Hill et al., 2011; Kane & Staiger, 2008). VAMs seek to estimate teachers' contribution to current learning, taking prior learning into account. Thus, using only direct measures of student prior learning, such as test scores, might be preferable and avoid bias because students' expected learning is not based on student characteristics and because student background characteristics are only proxy measures of prior learning (Ballou et al., 2004). However, test scores from multiple prior grades are not always available for all students, which might limit not just the number but also the kinds of students who are included in VAMs, potentially biasing those estimates. For example, student score histories might be less complete in urban districts with higher student mobility. Knowing that lack of a complete score history for most students can be compensated for by using student background characteristics as statistical controls might be useful because it makes VAEs within the same district more comparable. It also allows using larger numbers of students per teacher in the model, another important factor affecting VAE precision. That is why it seems important to understand whether the two approaches are comparable from a practical perspective in reducing variance in student learning that is unrelated to teacher performance.

A study by Ballou and colleagues (Ballou, et al., 2004) that investigated the effects of including student and school characteristics in addition to test scores in the TVAAS model found negligible effects on teacher VAEs classification into performance groups when student characteristics were included in addition to student test scores from multiple prior

years. The results suggest that test scores controlled sufficiently for differences in student prior learning. The study, however, did not investigate whether replacing multiple prior years of student test scores with student background variables provided comparable statistical control. When school characteristics were included in the TVAAS model, the percentage of reclassification of teachers into performance groups was more pronounced (Ballou et al., 2004), although the authors concluded that those results might be a consequence of estimation problems, rather than indicate the presence of school effects. Interestingly, VAMs tend to do little in the way they address school effects. Often school effects are difficult to control statistically because school districts do not collect systematic measures of school functioning (Raudenbush, 2004). In addition, school effects tend to be small compared to student and potential teacher effects and might be considered negligible. Nevertheless, between-school variation that is not accounted for in the model remains and is confounded in teacher VAEs, which is important especially when school districts are small and quite diverse. Understanding the effects of different approaches to statistical control is important for data considerations and interpretation of VAEs.

VAE Stability Over Time and Single Cohort Versus Multiple-Cohort VAEs

From a policy perspective, evaluating teacher performance on a yearly basis might be the most straightforward approach, although a number of studies have shown that teacher performance as measured by VAEs can vary considerably from year to year (Goldhaber & Hansen, 2010; McCaffrey et. al., 2009). If evaluated yearly, teacher performance will be estimated based on the difference between expected and actual learning of the cohort of students a teacher had in a given year (Ferrão, 2011; Measures of Effective Teaching Project, 2010). In fact, policies of many states that have currently implemented some form of value-added measures specify how many years of improved performance will be required to cancel out insufficient performance in prior years (National Council on Teacher Quality, 2011). Such policies assume not only that the required improved performance is possible (some states appear to mandate performance increases of up to two standard deviations for the weakest teachers in value-added distributions), but it also implies that value-added scores reflect, with some reasonable year-to-year fluctuation, a stable attribute (similar to expertise) that changes over time in some systematic way. In this view, extreme changes in VAEs from year to year for the same teacher, e.g., being in the lowest quartile in one year and in the highest quartile in the next, have been taken as an indication that VAEs contain substantial error and little signal of actual teacher performance. Others have pointed out that substantial differences in performance from one year to the next have also been documented for other fields, such as sales, and that teacher VAEs are fairly comparable to those reported performance fluctuations (Goldhaber & Hansen, 2010; McCaffrey et. al., 2009). In contrast, evidence for the stability of VAEs over time have been taken as an indication of more reliable estimates of teacher performance, although some have suggested that greater stability could also indicate bias in VAEs (McCaffrey et al., 2004). These are the primary reasons why some studies have examined the inter-temporal stability of teacher VAEs.

Although the number of studies that have investigated changes in VAEs over time is small and results differ somewhat by study, VAEs appear to be at best moderately stable with slightly more stability at the top of the performance distribution than at the low end. Studies report year-to-year VAE correlations ranging from 0.2 to 0.6, indicating that in some cases rank-ordering of teachers changed substantially (Buddin, 2010; Goldhaber & Hansen, 2010; McCaffrey et al., 2009). Similarly, studies that describe stability over time by classifying teachers into performance groups found that between one quarter to one half of the teachers

in the top quintile or quartile in one year remained in their group also the following year, while between 10 to 30% of the teachers fell into the lowest performance group in the second year (Aaronson, Barrow, & Sanders, 2007; Ballou, 2005; Goldhaber & Hansen 2010; Koedel & Betts 2007; McCaffrey et al., 2009). A comparable trend was observed for teachers who initially were in the bottom quartile (or quintile). Up to 40 percent of these teachers remained in their respective performance group also in year 2, while between 10 to 30% of the teachers showed up in the top performance group (Goldhaber & Hansen 2010).

Yet, still little is known about the determinants of within-teacher variation. In a variance decomposition analysis, McCaffrey and colleagues (2009) estimated that between 30 to 60% of the variance in teacher VAEs from year to year was due to sampling error from "noise" in student scores, whereas persistent teacher effects (not due to noise) accounted for 50% of the variation in VAEs for elementary school teachers and 70% for middle school teachers. McCaffrey and colleagues concluded that much of variation in VAEs is due to within-student variability, part of which is due to regression to the mean: Students who obtained very high scores on the standardized test in one year, are likely to obtain somewhat lower scores in the following year due to regression to the mean. Because those scores are modeled to estimate VAEs, some of that within-student variability will be reflected in VAEs.

As an alternative to less stable single cohort VAEs, one could analyze data from multiple student cohorts for a given teacher, if available. Such multiple cohort models estimate an overall, assumed-to-be-stable teacher effect for the years under study that takes within-teacher variation into account (Kane & Staiger, 2008). Because more students are included in the analysis per teacher, standard errors will be smaller than those obtained from a single cohort analysis, which improves precision and might affect teacher classification into performance groups. Ballou (2005) shows that precision of VAEs increase with the number of annual observations per teacher. Estimating teacher effects over a three-year span, 58% of middle school math teachers had estimates that are significantly different from the average teacher effect, whereas single-year estimates only designate 30 percent of teachers as being significantly different than the average. These findings have implications for policy makers because they need to decide what percentage of the population policy needs to cover. Can a policy that rewards the top 15% and sanctions the bottom 15% be sufficiently effective?

The differences between single and multiple cohort VAEs also raise another, related question, which directly ties into policy and how such scores might be used. To be most useful, policy and VAEs need to be compatible. For example, value-added scores based on multiple cohorts might be used when the stable component of estimated teacher performance is the focus of educational policy. If policy makers intend to reward teachers who, on average, perform well, as estimated by VAMs, or sanction teachers who, on average, perform way below average, multiple cohort VAEs are preferable over single cohort ones. However, such a policy would require that teachers have taught at least a specified minimum number of student cohorts in a given district or state to be evaluated and that teacher evaluations would occur only every few years. If on the other hand, the focus is on rewarding the higher performing teachers in any given year, single cohort VAEs might be more appropriate although they are less stable.

Student Sample Size in VAE Estimation

One criterion for the quality of any measure, including VAEs, is how well it helps to distinguish between members of its target population. The higher the precision of each individual person's estimate, that is, the smaller the associated standard error, the more

highly we rate the quality of a measure. In a typical measurement context, precision is most strongly influenced by the number and quality of items that are close to a person's "true" although unknown trait or ability (Embretson & Reise, 2000). In the value-added context, the number of students and the accuracy and precision of their scores directly influence accuracy and precision of teacher VAEs. A larger number of students with highly reliable information will produce more precise teacher VAEs than a smaller number of students with less reliable information. In general, estimates based on small samples tend to be less reliable and precise than those based on larger samples because in small samples there are fewer observations and the particular observations included can greatly affect the overall estimate (Cohen, 1988; Embretson & Reise, 2000). Outlying cases in small samples can pull sample estimates, such as student learning, considerably up or down, affecting either expected or observed learning and hence will influence the difference between those. Beyond the size of the standard errors, however, more students also mean a more accurate estimate of overall classroom learning. Systematically missing data for particular student subgroups within a classroom might bias teacher VAEs. We focused on the effects of student sample size on the precision of VAEs and potential changes in teacher performance classification.

Larger sample sizes result in smaller standard errors, which translate into smaller confidence intervals, and indicate greater precision. If teacher performance group designation considers VAE precision (Ballou, 2005; Ballou et al., 2004) as we did in this study, changes in student sample size will affect teacher performance. If a VAE is close to an adjacent performance group, a small confidence interval might identify the estimate clearly as belonging to that performance group while a larger confidence interval would indicate that the estimate could also have fallen into the adjacent group.

Although the overall number of students included in VAMs tends to be very high (often multiple hundreds of thousands), the number of students used in the estimation of each teacher's value added score can be surprisingly small. For any given cohort in elementary school, the number of students available for estimating each teacher's VAE might range from 20 to 30 depending on class size, but actual sample sizes of 10 to 15 students per teacher or even less in a given year are not uncommon due to missing data (Mc Caffrey et al., 2009). Cohort numbers for middle school and high school teachers tend to be higher because they teach multiple courses in their subject area, although often such models include course information (McCaffrey et al., 2009). To improve precision of VAEs, some studies specify a minimum student sample size (Buddin, 2010). Goldhaber and Hansen (2010) only include teachers who have at least 10 students with available data and no more than 29 to exclude teachers who might face unusual teaching situations such as team teaching.

In many cases minimum sample sizes around 10 to 15 students are specified (Buddin, 2010; Goldhaber & Hansen, 2010; McCaffrey et Al., 2009) although rationales for such determinations differ. Some have estimated VAEs based on a minimum of five students per teacher (Harris & Sass, 2010). One approach to determine the minimum number of students required is based on the size of standard errors. For example, McCaffrey and colleagues (2009) who systematically evaluated changes in standard errors as a function of student sample size, found that mean standard errors were five to six times as large when value-added scores were estimated based on 1-4 versus more than 20 or more students, and that student samples sizes of smaller than 15 students introduced considerable imprecision into teacher VAEs. They observed that when scores are normed to have a mean of zero and a standard deviation of 1, a standard error of .5 indicates that even if teachers had no true effects and all the variability among teachers was due to sampling error,

the variability among teachers would equal about 25 percent of the variance among students. Using standard errors to specify desired levels of estimate precision has a long tradition in psychometrics. Standard errors around .3 indicate that estimates are considered highly reliable ($r_{eliability}$=.90) (Embretson & Reise, 2000). Different teachers in the same jurisdiction (i.e., district or state) will have different numbers of students for whom test scores are available, which might lead some teachers to be classified into a different performance group than they might have had they had more student data.

Methods

Data and Sample Description

Data from four sequential cohorts (2004-2007) of 5th grade students and their teachers in a large urban district in California were made available by the district administration for analysis. For each cohort of students, 3rd grade, 4th grade, and 5th grade achievement scores (IRT scaled scores) were available for the mathematics and English Language Arts (ELAs) sections of the state standardized test. Our initial data set contained information for 3878 5th grade mathematics teachers and a total of 208,137 students in 474 schools. We removed all students for whom we did not have complete mathematics and ELA standardized test scores for grades 3, 4, and 5 and all relevant student background information we intended to include in the models. We also removed all students and teachers who had missing identification (i.e., no teacher, student, or school ID), and teachers who changed schools during the years we analyzed. The goal was to create a single, complete, and clean data set for all analyses that would allow attributing any changes in teacher rankings and performance group classifications to particular changes in model and data specifications. The exclusions left us with a data set that included information for 3651 fifth grade mathematics teachers and their 161,811 students in 469 schools for the four years of study. The breakdown by cohort and the combined total is summarized in Table 1. Note that the combined total information reported in Table 1 is not the sum of the individual cohorts because many teachers have taught multiple cohorts of students. It represents the total number of 5th grade teachers, students, and schools in the district during the four years of study for which complete data were available. Interestingly, for almost half of the 3561 teachers only a single cohort of fifth grade students was available (45.2%; N = 1652). For about a quarter of the teachers' data from two 5th grade student cohorts were available (23%; N = 841). Roughly half of the remaining teachers had data from three 5th grade student cohorts (14%; N = 510) and the other half had four cohorts (17.7%; N = 648). Although this distribution might suggest that the district experienced significant teacher turnover during the years of the study, another explanation could be that a considerable number of our 5th grade teachers taught at different grade levels in prior years. Our analyses focused exclusively on 5th grade mathematics teachers, and, therefore, student data from teaching other grade levels were not included, although some models (e.g., EVAAS) run these calculations on teachers who switch grades. Nevertheless, it is important to note for practical use of teacher VAEs that even in a teacher sample of this size, nearly one half of the teachers (45.2%) have student data from only a single 5th grade student cohort.

Table 1

Sample Structure by Cohort and Combined

Year (Cohort)	2004	2005	2006	2007	Total (combined)
Original Number of students	53,442	54,402	51,735	48,558	208,137
Original Number of teachers	2,042	2,119	1,926	1,891	3,878
Original Number of schools	446	454	462	473	474
Number of students used in analyses	39,741	42,715	40,852	38,503	161,811
Number of teachers used in analyses	1,936	1,949	1,843	1728	3,651
Number of schools used in analyses	444	447	454	467	469

The following variables were used in the analyses reported in this paper: Student mathematics and ELA scaled scores (grades 3-5) from the standardized state test for school years 2003/04 (cohort 1) through 2006/07 (cohort 4), student gifted status, free or reduced lunch as a proxy for SES, special education status, and ethnicity. Table 2 shows this information for the analyzed sample by cohort and overall. Note, that we included students' scaled scores and corresponding z-scores (z-scores are explained below) in Table 2 because the scaled scores are used to describe student performance groups with regard to the grade level standards. The mathematics and ELA sections of the standardized test have a score range from 150 to 600. For example, scores between 300 and 349 indicate that students achieve basic performance on the grade level curriculum standards, while scores between 350 and 399 reflect proficient performance. We rescaled student scores by grade level to have a mean of 0 and a standard deviation of 1, and used these z-scores in our statistical analyses because tests were not vertically equated across grade levels and because they provide a metric for VAEs and their associated standard errors that would be easy to interpret (McCaffrey at al., 2009). The remaining variables had either a value of 1 or 0; 1 indicating that a student is classified as gifted, a recipient of free or reduced lunch, eligible for special education services, or that a student identified with a particular race/ethnicity, and 0 otherwise. We also computed the mean mathematics and ELA achievement for each classroom and school, and the percentage of students who are classified as gifted, receiving special education service, and free or reduced lunch to account for potential classroom and school factors that might affect student learning but over which teachers might have little or no control. It is important to note that the measures of potential classroom and school effects we created by aggregating student information may not be of sufficient quality to account for actual classroom and school effects should such effects exist (Raudenbush, 2004). Rather, much of the variation such aggregate variables are likely to account for, will have been controlled for by the student level information. Indeed, preliminary analyses

revealed that school-level inputs did not explain any additional variance in student scores, while some classroom level inputs did. As a consequence we excluded all school level inputs from our final models. Descriptive statistics for all variables included in our final models are presented in Table 2.

Table 2 shows that, overall, student performance on the state test both in mathematics and ELA reflects basic performance (scores range between 300 and 349). Under basic performance, students demonstrate a partial and rudimentary understanding of the mathematics and ELA knowledge and skills required for 5th grade. Math performance is slightly higher than ELA performance, perhaps a function of the high percentage of students in the district who are English language learners (ELLs). Overall, student performance in the district increased slightly over four years both in mathematics and ELA, a trend that might be interesting to policy makers and analysts but that is not captured in VAEs. VAEs compare teacher performance to the district's average performance in any given year. Although a baseline performance could be specified in VAMs so that teacher performance during later years can be estimated against that baseline, such models have not yet been used much in practice.

Table 2 also shows that the percent of students identified as gifted has increased slightly over the four years, from 14% in 2004 to 20% in 2007. The percentage of students eligible for special education has remained virtually unchanged (6% to 7%) over the same time period, while the percentage of students receiving free or reduced lunch has slightly decreased (from 89% to 81%), but still appears high, which is not uncommon for large urban districts. If students were randomly assigned to teachers, one might imagine an average 2007 classroom in this district as consisting of about 22 students, of which 18 students receive free/reduced lunch, about 2 students might receive special education services, and 4 students might be classified as gifted, which appears to put the gifted distinction at about 1 standard deviation above the mean. To be sure, there is evidence to suggest that students are not randomly assigned but sorted by ability or other criteria (Rothstein, 2007).

One important question to ask might be if the particular characteristics of this student sample might have any bearing on the results of this study. Are results likely to be substantially different if, for example, the percentage of ELLs would have been larger or the percentage of students receiving free or reduced lunch would have been smaller? While our sample clearly reflects characteristics of an urban school district, these are important questions when considering how generalizable results from this study might be beyond this particular district.

Table 2
Student, and Teacher Characteristics by Cohort and Combined

Variables	Cohort 1 2003/04		Cohort 2 2004/05		Cohort 3 2005/06		Cohort 4 2006/07		Overall	
Student-level	N=39,741		N=42,715		N=40,852		N=38,503		N=161,811	
	M	SD	M	SD	M	SD	M	SD	M	SD
MATH3CST	319.68	67.99	337.9	72.41	348.58	74.93	358.17	75.06	340.94	74
MATH4CST	341.73	67.19	339.45	63.77	347.6	70.06	355.26	72.73	345.83	68.68
MATH5CST	332.37	71.55	341.73	84.47	346.33	85.91	351.64	83.54	342.95	81.94
ZMATH3	0.01	0.98	0.03	0.98	0.02	0.99	0.03	0.98	0.02	0.98
ZMATH4	-0.02	0.99	0.01	0.99	0.01	0.99	0.02	0.99	0.01	0.99
ZMATH5	-0.03	0.99	0.01	0.99	0.01	1	0.02	1	0	1
ELA3CSTS	309.99	54.7	311.87	53.76	308.81	53.39	311.41	54.57	310.53	54.11
ELA4CSTS	327.3	44.01	325.98	45.8	332.21	49.82	338.05	54.66	330.75	48.88
ELA5CSTS	327.14	47.44	327.36	48.23	330.03	50.84	332.1	49.45	329.11	49.04
ZELA3	-0.01	0.98	0.01	0.97	0	0.98	0.01	0.98	0	0.98
ZELA4	-0.04	0.99	-0.01	0.97	0	0.98	0.01	0.98	-0.01	0.98
ZELA5	-0.05	0.99	0	0.97	0.01	0.98	0.01	0.99	-0.01	0.98
GIFTED	0.14	0.35	0.15	0.36	0.18	0.38	0.2	0.25	0.17	0.37
SPED	0.06	0.23	0.07	0.25	0.07	0.26	0.07	0.26	0.06	0.25
REDUCEM	0.89	0.31	0.89	0.32	0.83	0.37	0.83	0.38	0.86	0.35
R1 (Hispanic)	0.79	0.41	0.78	0.41	0.79	0.41	0.78	0.42	0.78	0.41
R2 (Pacific Islander)	0	0.05	0	0.06	0	0.05	0	0.05	0	0.05
R3 (Black)	0.09	0.29	0.09	0.29	0.09	0.29	0.09	0.29	0.09	0.29
R4 (Filipino)	0.02	0.13	0.02	0.14	0.02	0.14	0.02	0.15	0.02	0.14
R5 (Asian)	0.03	0.18	0.03	0.18	0.03	0.18	0.03	0.18	0.03	0.18
R6 (American Indian)	0	0.04	0	0.05	0	0.05	0	0.05	0	0.05
Teacher-level	N=1936		N=1949		N=1843		N=1728		N=3651	
Mean by T	M	SD	M	SD	M	SD	M	SD	M	SD
MATH3CST	319.68	44.37	337.69	46.32	348.09	47.28	356.65	47.55	336.04	43.96
MATH4CST	341.68	42.79	339.36	42.12	347.09	44.19	353.79	45.81	341.59	40.33
MATH5CST	332.34	49.93	341.53	57.64	345.33	57.87	349.78	55.56	336.5	51.65
ZMATH3	0.01	0.64	0.02	0.62	0.01	0.62	0.01	0.62	-0.04	0.58
ZMATH4	-0.02	0.63	0.01	0.65	0.01	0.63	0	0.63	-0.06	0.58
ZMATH5	-0.03	0.69	0	0.68	0	0.67	0	0.66	-0.07	0.63
ELA3CST	310.31	38.07	311.98	36.62	308.69	35.86	310.78	37.11	307.27	34.08
ELA4CST	327.65	30.32	326.09	31.53	332.15	34.03	337.09	37.14	327.78	30.86
ELA5CST	327.47	32.34	327.44	33.52	329.74	34.3	331.22	33.44	325.79	31
ZELA3	0	0.68	0.01	0.66	0	0.66	0	0.67	-0.06	0.62
ZELA4	-0.03	0.68	-0.01	0.67	0	0.67	-0.01	0.67	-0.07	0.62
ZELA5	-0.04	0.68	0	0.67	0	0.66	-0.01	0.67	-0.08	0.62
GIFTED	0.16	0.27	0.16	0.26	0.19	0.26	0.21	0.27	0.16	0.24
SPED	0.07	0.12	0.08	0.12	0.08	0.12	0.09	0.15	0.08	0.13
REDUCEM	0.87	0.25	0.87	0.24	0.82	0.25	0.81	0.25	0.85	0.23
R1 (Hispanic)	0.76	0.29	0.76	0.29	0.76	0.28	0.75	0.29	0.77	0.27
R2 (Pacific Islander)	0	0.02	0	0.02	0	0.02	0	0.01	0	0.01
R3 (Black)	0.10	0.19	0.10	0.19	0.10	0.18	0.10	0.19	0.10	0.18
R4 (Filipino)	0.02	0.06	0.02	0.05	0.02	0.05	0.02	0.06	0.02	0.05
R5 (Asian)	0.04	0.10	0.03	0.10	0.04	0.10	0.04	0.10	0.03	0.08
R6(American Indian)	0	0.02	0	0.01	0	0.01	0	0.01	0	0.01

Some characteristics of urban districts were also reflected in our sample of 5th grade teachers. Teachers' reported ethnicity was quite diverse. For those teachers (N=3505) who reported their ethnicity, less than half (42%) identified as white/non-Hispanic, while 33% identified as Hispanic, 14% as African-American, 11% as Asian or Pacific Islander, and less than one percent (0.4%) as American-Indian or Alaska Native.

In other respects, the sample of teachers was quite typical for elementary grades. With regard to the 3505 5th grade mathematics teachers in the district for whom information was available (out of total N = 3651), more than two thirds (70%) were female. About three quarters (N = 3502) reported that they had earned either a bachelor degree (33%) or a bachelor degree with 30 additional semester hours (41%). Ten percent of the teachers reported holding a masters degree, 15% reported a master's degree plus 30 semester hours, and 1% of teachers had earned a doctoral degree.

The teachers reported an average of 10.7 years of teaching experience (SD=8.6) with a minimum of 1 year and a maximum of 50 years (N = 3503). Nearly all teachers with information (94% of N = 3495) were fully credentialed; the remaining teachers either were in the process of obtaining a credential or held a preliminary credential. Note, that the description reflects 2007 information for all time-varying teacher characteristics.

Estimating Teacher Effects based on Value-added Approaches

We estimated teacher effects from 2-level and 3-level mixed models. We fit 2-level models to estimate teacher VAEs based on single cohorts of students. We fit a 3-level, multiple-cohort model to estimate VAEs representing teachers' weighted average performance as described by Kane & Staiger (2008). Prior to analyses we centered our independent variables around the grand mean. For simplicity, in both model notations below, *S* represents a vector of student level inputs, such as scores from prior grades and student characteristics, for example, participation in a gifted program, etc.; *C* represents a vector of student specific classroom-based inputs, such as percent of students who participate in a gifted program in the class.

2-Level Model

In the 2-level model, students (level 1) are nested within teachers (level 2). There was no need to model different classes for teachers, because in 5th grade classrooms are still self-contained. In the basic, single cohort, 2-level mixed model we can denote student achievement level in grade 5 as

$$\text{Math Grade } 5_{ij} = \square_{00} + \square_{01}(C) + \square_{10}(S) + \square_{0j} + \square_{ij} \quad \text{[Eq. 1]}$$

The 5th grade achievement level of student i of teacher j is a function of all student level (S) and all classroom level inputs (C). Student level inputs are included at level 1, classroom level inputs are included at level 2. The term $\square_{ij}$ is a residual term that captures between student differences within teachers (at level 1). The level 2 residual, denoted as $\square_{0j}$ represents the difference between the expected and observed learning for each teacher as a deviation from the grand mean (i.e., district mean). This term contains the VAEs, which are presumed to represent teachers' contribution to student learning. Note, that because no school effects are included in the model, the teacher effect is measured relative to the average of all 5th grade mathematics teachers in the district. It is important to mention that school effects will be confounded in teacher VAEs, unless they can be sufficiently controlled

for in the model. This might be less of a concern as long as school effects are small, but it might be of concern if school effects are large or can't be sufficiently controlled for.

Multiple-cohort Model

In the multiple-cohort model, students (level 1) are nested within cohorts (level 2), and cohorts are nested within teachers (level 3). Kane and Staiger (2008) described this model in detail. Under this model a single VAE is estimated for each teacher that is assumed to be stable over the years under study (level 3), after controlling for within-teacher variation. In addition, the model estimates n cohort-specific residuals at level 2, which represent each teacher's yearly deviation from his or her estimated stable VAE.

In the basic, multiple cohort model we can denote student achievement level in grade 5 as

$$\text{Math Grade } 5_{ijk} = \pi_{000} + \pi_{001}(C) + \pi_{100}(S) + u_{00k} + r_{0jk} + e_{ijk} \quad \text{[Eq. 2]}$$

In the multiple cohort model, the 5th grade math achievement level of student i in cohort j of teacher k is a function of all student level (S) and all classroom level inputs (C). In this model, student level inputs are included at level 1 and classroom and school level inputs are included at level 3. The term e_{ijk} is a residual term that captures student differences within cohorts within teachers (at level 1); the residual term r_{0jk} captures within-teacher variation across cohorts, and u_{00k} represents the between-teacher variation around the district mean π_{000}. Again, no school effects are included, and the stable teacher effect is measured relative to the weighted average of all 5th grade teachers in the district.

Specific Models by Research Question

Below, we describe for each research question, which specific models were fit. Table 3 summarizes the different models by research question.

Research Question 1:

What are the effects of statistical controls on the stability of teacher VAEs? To answer this research question, we fit four 2-level, single cohort, models to the 2007 cohort data (denoted Model 1 through 4 in Table 3). We varied the statistical controls included (previous year's test scores only versus test scores from two prior years, and student background characteristics), and we also compared two different strategies for controlling, either using test scores only, or a combination of test scores and other student characteristics known to affect learning. By specifying these particular models we were able to explore (a) how much variance in student test scores we were able to control for overall in the four models and (b) how different ways of controlling for student differences affected teacher VAEs.

Research Question 2:

How stable are single cohort VAEs over time and how do they compare to stable VAEs estimated from multiple student cohorts? To address this research question, we fit Model 1, in which we controlled only with student test scores from the previous year to each cohort, obtaining four separate cohort VAEs. We used these single cohort VAEs to evaluate stability of VAEs over time for teachers (N = 648) with four student cohorts. We also fit a multiple cohort model using all four cohorts (Model 5 in Table 3), in which we again controlled only for previous year test scores. We compared VAEs from each single cohort with the stable VAE obtained from the multiple cohort model.

Research Question 3:

What are the effects of student sample size on the precision of teacher VAEs and performance group designation? To explore this research question we combined student data from all four cohorts and analyzed the data as a single cohort to maximize each teacher's student sample size. We fit Model 1 to these combined data. We then determined the average standard error associated with VAEs based on 10, 20, 30, 40, and 50 students and used those averaged standard errors to compute five confidence intervals around each teacher's VAE. We chose to specify student sample sizes rather than number of cohorts because the number of students per teacher in a given cohort might vary substantially. By keeping teacher VAEs constant and using different size standard errors, we focused exclusively on sample size effects with regard to the precision of the estimates.

Table 3
Model descriptions by Research Question

Model	Research Question 1 for 2007 cohort only	Research Question 2 For all four cohorts	Research Question 3 For all four cohorts combined
(1) 2-level single-cohort model: Level 1: grade 4 Math & ELA Level 2: grade 4 mean Math & ELA	X	X	
(2) 2-level single-cohort model: Level 1: grade 4, 3 Math & ELA Level 2: grade 4, 3 mean Math & ELA	X		X
(3) 2-level single-cohort model: Level 1: grade 4 Math & ELA, Gifted, Special Ed, SES, & Ethnicity Level 2: grade 4 mean Math & ELA, percent Gifted, Special Ed, & SES	X		
(4) 2-level single-cohort model: Level 1: grade 4, 3 Math & ELA, Gifted, Special Ed, SES, & Ethnicity Level 2: grade 4, 3 mean Math & ELA, percent Gifted, Special Ed, & SES	X		
(5) 3-level multiple-cohort model: Level 1: grade 4 Math & ELA Level 2: - Level 3: grade 4 mean Math & ELA		X	

Results

The results are organized into four sections. The first section reports on how variance in student scores is distributed in our data. Each subsequent section presents results pertaining to one of the three research questions.

Variance in Student Scores

An important first step in our value-added analyses was to estimate how the variance in student scores, that is, grade 5 mathematics scores, was distributed among students within teachers, among teachers within schools, and among schools, when no control variables were included (unrestricted model). If, for example, the variance in student scores that is estimated to be between teachers represents a very small percentage of the overall variance in student scores, teacher effects which are assumed to be the source of this between-teacher

variance would be small, and differences in student learning would depend little on which teacher a student had. In such instances, value added measures might not provide a meaningful distinction of teacher performance. If, on the other hand, a considerable amount of variance in student scores falls between teachers, proponents of value-added approaches might conclude that learning outcomes of students appear to depend to some degree on which teacher they had. How much of this between-teacher variation represents actual teacher effects or other systematic sources of variation, continues to be a topic of further study and debate.

When we partitioned the variance in students' 2007 5th grade mathematics scores between students and between teachers (2-level model), 62% of the total variance was between students and 38% was between teachers, suggesting that there is considerable between-teacher variation. It is important to note that under variance decomposition any variance between schools would be contained in the between-teacher variance because school effects are not included in this model. If school effects would be modeled, 61% of the variance would still be between students within teachers, 24% would be between teachers within schools, and 14% of the variance would be among schools. Variance components estimated for the remaining three cohorts (2004, 2005, and 2006) showed nearly identical distributions.

We also estimated how the variance in student scores would be partitioned under the multiple cohort model, in which students (level 1) are nested within cohorts (level 2), and cohorts are nested within teachers (level 3). Under this model, 61% of total variance in students' 5th grade scores was between students within cohorts, 7.5% among cohorts within teachers, and 32% was between teachers. No school effects were included in this model and any variation due to schools is part of the between-teacher variation.

Several observations are noteworthy from a value-added perspective: (1) A considerable amount of variance in student scores is estimated to fall between teachers, which suggests that if teachers are a large and systematic source of that variation and other sources of variation are comparatively small, differences in estimated teacher performance will be meaningful. The breakdown of variance in student scores in our study is similar to percentages reported in other studies (Rowan, Miller, & Correnti, 2002; Hatti, 2003). (2) About 14% of the estimated between-teacher variance (38%), is associated with between-school variance, when school effects are modeled, which means that between-school variation is not trivial and might be expected in such a large and diverse urban district. Nevertheless, there is almost twice as much variance between teachers within schools (24%) than across schools (14%), indicating that differences between teachers within the same school are more pronounced than differences between schools. In VAMs that do not include school effects, like the ones we used in this study, this variance will be part of the teacher VAEs, unless they can be sufficiently controlled for. (3) There is comparatively little within-teacher variation across cohorts (7.5%), suggesting some consistency across cohorts. This summarizes the variation and the sources of variation in student scores, which will be modeled to estimate teacher performance.

Research Question 1: What are the effects of statistical controls on the stability of teacher VAEs?

To answer this question, we fit Models 1 through 4 to the 2007 cohort data. In two of the models (Models 1 & 2) we only controlled for student prior test scores (previous year only, and two prior years), and in the other two models we controlled for a combination of student test scores and student background characteristics (previous year test scores plus

student background characteristics, Model 3, and two prior years test scores plus student background characteristics, Model 4). In all four models we included classroom-level aggregates of the respective student variables (at level 2) to account for classroom effects that lie outside the teacher's control. Table 4 summarizes the results from the four models.

Table 4
2007 Results: 2-level Model

Fixed Effects	Null	Model 1	Model 2	Model 3	Model 4
Intercept	0.01	-0.00	0.01	0.01	0.01
(SE)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
p if not <0.05	*0.74*	*0.95*	*0.07*	*0.14*	*0.10*
Math Grade 4		0.60**	0.46**	0.56**	0.44**
		(0.01)	(0.01)	(0.01)	(0.01)
Math Grade 3			0.26**		0.24**
			(0.01)		(0.01)
ELA Grade 4		0.22**	0.15**	0.19**	0.13**
		(0.01)	(0.01)	(0.01)	(0.01)
ELA Grade 3			0.017**		0.01
			(0.006)		(0.01)
					0.29
Gifted Status				0.32**	0.21**
				(0.01)	(0.01)
Special Ed Status				-0.08**	-0.04**
				(0.01)	(0.01)
Free/reduced Lunch				-0.01	-0.01
				(0.01)	(0.01)
				0.25	*0.13*

Table 4. (Cont.'d)
2007 Results: 2-level Model

Fixed Effects	Null	Model 1	Model 2	Model 3	Model 4
Hispanic				-0.04*	-0.03
				(0.02)	(0.02)
					0.03
Pacific Islander				-0.03	-0.01
				(0.06)	(0.06)
				0.63	*0.90*
African American				-0.11**	-0.08**
				(0.02)	(0.02)
Filipino				0.11**	0.11**
				(0.02)	0.02
Asian				0.16**	0.15**
				(0.02)	(0.02)
American Indian				-0.03	-0.04
				(0.06)	(0.06)
				0.56	*0.51*
Mean Math Grade 4		0.04	0.10*	0.02	0.09*
		(0.04)	(0.04)	(0.04)	(0.04)
		0.36		*0.63*	
Mean Math Grade 3			-0.03		-0.08
			(0.05)		(0.05)
			0.58		*0.09*
Mean ELA Grade 4		0.06	-0.03	-0.03	-0.05
		(0.04)	(0.06)	(0.04)	(0.06)
		0.10	*0.59*	*0.48*	*0.33*
Mean ELA Grade 3			0.01		0.00
			(0.05)		(0.05)
			0.79		*0.94*
Percent Gifted				0.15*	0.18**
				(0.07)	(0.067)
Percent SpecEd				-0.10	-0.10
				(0.06)	(0.06)
				0.11	*0.09*
Percent Free/reduced Lunch				0.07	0.08
				(0.06)	(0.06)
				0.19	*0.18*
Random Effects					
Intercept u0	0.39**	0.091**	0.089**	0.087**	0.087**
Residual	0.63	0.24	0.21	0.23	0.21
% Variance Explained		68	70	69	72

*p<.05, **p<.01

Several observations are noteworthy. Overall, we were able to control for about 70% of total variance in student scores with little difference between the four models. We controlled for about two thirds of the variance between students and some variance between teachers and schools, all of which represents variation in student achievement that is assumed to be unrelated to teacher performance. Under these models, the remaining 30% of variance in student scores would be attributed to differences in teacher performance, quantified in VAEs. Perhaps surprisingly, the amount of variance reduction was similar between the model that least and most controlled for differences in students. Controlling only with previous year test scores accounted for nearly as much variance (68%) as controlling with test scores from two prior years and additional student characteristics

(72%). Further, Table 4 also shows that controlling with a combination of previous year test scores (grade 4 in mathematics and ELA) and other student background variables is comparable to models that use test scores from two prior years, which might be important for school districts that have higher student mobility and that might be less likely to have multiple years of tests scores for most students.

Given that the reduction in variance was similar across the four models, the different approaches to controlling for differences in students had little effect on teachers' relative standing. Pairwise correlations between teachers' VAEs from these four models were very high, ranging from .97 to .99, suggesting that rank ordering of teachers was very stable.

However, VAEs that are highly correlated can result in changes in teachers' performance group designation because even small changes in the value-added distribution might lead teachers to be designated into a different performance group, which is important for policy makers. If we assume that controlling for more variance in student test scores will reduce between-teacher variance, while standard errors remain the same, we can expect that performance group designation will change for some teachers (Ballou et al., 2004). Although the difference in reduction of variance is only 4% between Model 1 and Model 4, we computed the percentage of teachers who changed or remained in their respective performance group for all pairwise model comparisons. That is, the percentage of teachers with average performance as measured by their VAEs and associated standard errors and the percentage of teachers significantly above or below average performance.

Between 4 to 12% of teachers were reclassified under pairwise comparisons, depending on the compared models; between 88% and 94% of the teachers remained in their respective performance groups regardless of how we controlled in our models, which indicates a fair amount of stability. Because the pairwise model comparisons do not provide information about which teachers were reclassified, we also computed the percentage of teachers that were consistently classified into the same performance group under all four models. In total, 84% of all teachers remained stable in their performance group designation, while the remaining teachers (16%) showed unstable classification across the four models. About half of the teachers were consistently classified as average effective teachers (49.2), 35% of them were consistently classified as significantly different from average performance (17.4% significantly above average, and 17.4% significantly below average). Overall, changes in statistical controls led to little instability. In other words, missing 3rd grade scores or other student information for some teachers would have had little effect on teacher' VAEs and their performance group designation within the district, as long as previous test scores were available.

Research Question 2: How stable are single cohort VAEs over time and how do they compare to stable, multiple-cohort VAEs?

If teacher VAEs are presumed to reflect teacher performance, it is important to understand how stable teacher performance is over time. Those who think of teacher performance as a fairly stable attribute, with some year-to-year fluctuations, might interpret considerable changes in performance from one year to the next as an indication of rather noisy estimates. For this analysis we only selected those teachers in our sample for whom we had estimated VAEs (Model 1) for all four years under study (N = 648). Within-teacher, year-to-year correlations are shown in Table 5.

Table 5
Year-to-Year Correlations for VAEs from Model 1

Models	2007	2006	2005
Single cohort 2006	.66**	1	
Single cohort 2005	.60**	.65**	1
Single cohort 2004	.52**	.59**	.62**

N=648

Interestingly, year-to-year correlations between VAEs (Table 5) were fairly large and stable over time. Correlations between adjacent years ranged from .62 to .66. Correlations between non-adjacent years decreased as the time interval between cohort increased. Thus, over the four years of study, teacher performance might be considered moderately stable, perhaps even stable, and although rank ordering of teachers clearly changed, there is also evidence of a fair amount of consistency.

In terms of performance group designation, across adjacent years 70% of teachers were stable, and 30 percent changed performance groups. Even over the four years of study, 55% of teachers were still in the same performance group that had been in during Year 1, while 45% were in different performance groups. Clearly, VAEs are less stable over time than when we had varied statistical controls. When comparing adjacent years, twice as many teachers were reclassified as were based on varying statistical controls, across all four years three times as many teachers were reclassified.

Among the many possible explanations for changes in teachers' relative standing, real changes in teacher performance is one of them. To explore if the correlation patterns we observed might be an indication of systematic change, we divided the teachers in our sample into three groups. Group 1 consisted of teachers who had the least experience, between 1 and 5 years; group 2 included teachers with 6 to 10 years of experience; and group 3 consisted of the most experienced teachers, those with more than 10 years of experience. Year-to-year correlations by teacher experience are shown in Table 6.

Table 6
Year to Year Correlations between Single-cohort VAEs based on Model 1 by Teacher Experience

Less than 5 years (N=82)	2007	2006	2005	2004
2006	.65**	1		
2005	.45**	.56**	1	
2004	.28**	.39**	.39**	1
5 to 10 years (N=246)	**2007**	**2006**	**2005**	**2004**
2006	.69**	1		
2005	.62**	.64**	1	
2004	.55**	.57**	.61**	1
More than 10 years (N=320)	**2007**	**2006**	**2005**	**2004**
2006	.63**	1		
2005	.60**	.66**	1	
2004	.55**	.64**	.66**	1

The correlation pattern shown in Table 6 suggests systematic changes in VAEs by teacher experience. For teachers with the least experience (1 through 5 years), correlations between VAEs estimated from the first and last cohort were half (.28) of the original size

(.52) reported for the entire sample, while correlations for teachers in the other two experience groups were stable and moderate to large. A repeated measures analysis that modeled yearly teacher VAEs by experience groups showed that overall teacher performance as measured by VAEs did not statistically significantly change over the four years of study ($F(3,1935)= 2.34$, $p>.05$). However, the interaction between VAEs and different experience groups was statistically significant ($F(6,1935)=2.93$, $p<.05$), indicating that teacher groups with different levels of experience had different VAE trajectories. Perhaps not surprisingly mean VAEs of teachers with the least experiences (<= 5 years) increased significantly from cohort 1 to cohort 4, while mean VAEs in the other two groups did not.

Next we compared the single cohort VAEs to the stable multiple cohort VAE, which is assumed to reflect teacher average performance over a specified time period, to further explore stability. Correlations between the yearly estimates and the stable multiple-cohort estimate were high, .88 or .89, depending on cohort. These findings indicate that teacher rank-ordering in each of the four cohorts was highly stable when compared with rank-ordering based on teacher average performance as estimated from multiple cohorts, and that each cohort contributed somewhat equally to the multiple cohort estimates. Nevertheless, changes in teacher rank-ordering between single and multiple cohort estimates were more pronounced than those observed when examining effects of statistical controls on VAEs. There, we reported correlations as high as .97 and .99.

Again, we examined how teacher classification into performance groups changed under the single and multiple-cohort models. When comparing any of the four, single cohort VAEs to the stable multiple cohort VAEs, between 26 and 28% of teachers were reclassified into a different performance group, while more than two thirds of teachers remained stable in their respective performance groups. Of those, slightly more than half of the teachers (between 53% and 54%) were consistently classified as performing average, while a total 19% of teachers was consistently classified as either above or below average performance. Consistent with the changes in teacher rank-ordering, reclassification percentages were less pronounced than we observed for teachers' yearly estimates, but more pronounced than we observed for statistical controls.

Research Question 3: What are the effects of student sample size on the precision of teacher VAEs and performance group designations?

To maximize student sample size for this analysis, we combined all four cohorts into one, and analyzed the data as one single cohort. We used teacher VAEs estimated from Model 2. Using these VAEs we computed five, different size standard errors by averaging the estimated standard errors of those teachers who had data from exactly 10, 20, 30, 40 or 50 students, respectively. We then computed for each VAE five different confidence intervals using those five different standard errors. We did not include VAEs for teachers with fewer than 10 students. Table 7 shows the changes in size of standard errors as a function of number of students used in estimation.

Table 7
Mean Standard Errors for Teacher VAEs by Number of Students

Number of Students	SE
10	.13
20	.10
30	.084
40	.073
50	.066

Table 7 shows that the size of the standard error associated with teacher VAEs based on 10 students decreases by about half when 50 students are assumed, suggesting that estimates based on the larger number of students will be twice as precise. The pattern of decreasing standard errors is consistent with what might be expected given that standard errors are a function of the square root of the sample size. Although sample sizes of more than 30 students are not likely to occur in any single cohort, the increase of precision between 10 and 30 students is still noticeable. Figures 1 through 5 depict the teacher VAEs and their associated confidence intervals based on variation in student sample size. As can be clearly seen, the percentage of teachers with confidence intervals that do not include zero (i.e., teachers who are significantly different from the mean) increases considerably as sample size increases. In other words, the number of teachers classified as below average and above average effective increases as the number of students on whose performance the estimate is based increases. Figure 5 shows three teacher performance groups of nearly equal size.

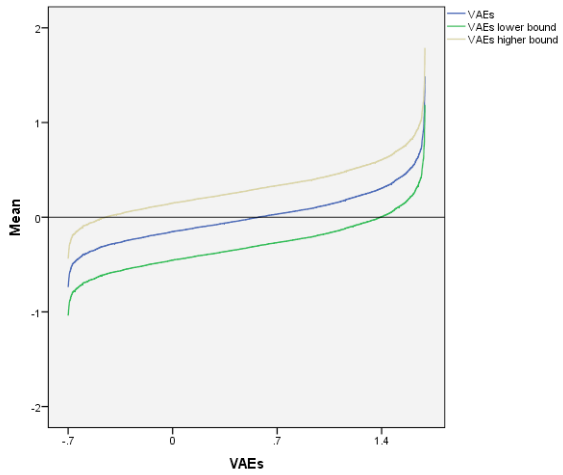

Figure 1. VAEs with Confidence Intervals for 10 Students

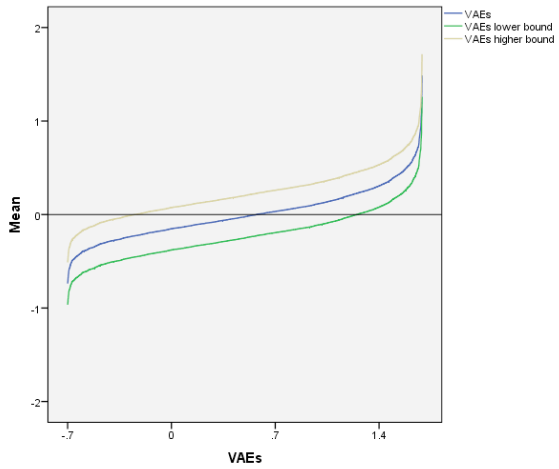

Figure 2. VAEs with Confidence Intervals for 20 Students

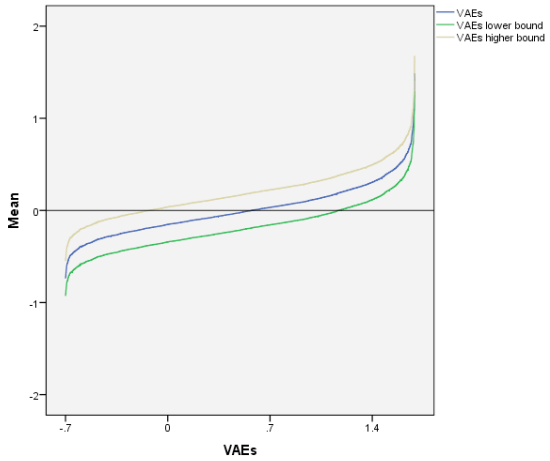

Figure 3. VAEs with Confidence Intervals for 30 Students

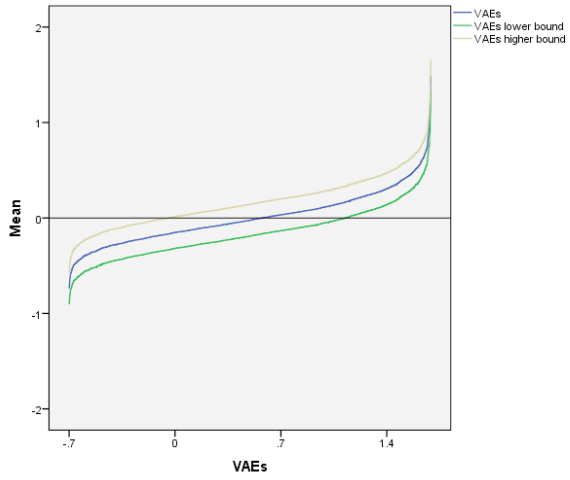

Figure 4. VAEs with Confidence Intervals for 40 Students

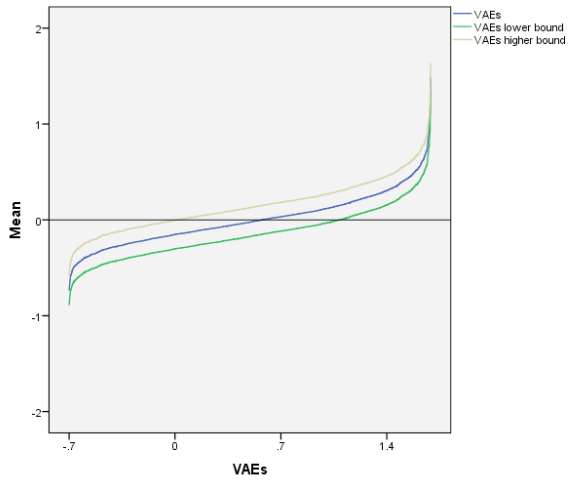

Figure 5. VAEs with Confidence Intervals for 50 Students

Again, we measured the impact of student sample size on teacher performance group designation. It is important to recall that reported changes only reflect changes in VAE

precision as described by different size standard errors. First, the number of teachers classified as being different from average effectiveness increases as sample size increases because standard errors decrease. For 10 students, 24% of VAEs are significantly different from average, for 20, 30, 40 and 50 students those values are 37%, 46%, 52%, and 55%, respectively. This was also observable in Figures 1 through 5. Second, the percentage of reclassified teachers based on the largest difference in sample size (10 versus 50 students) is 32 percent, which indicates that sample size had the largest relative effect on VAEs among the three model specifications. Reclassification percentages for the remaining comparisons were 28% (10 versus 40 students), 23% (10 versus 30 students), and 14% for 10 versus 20 students. Again, we computed the percentage of teachers who remained in their respective performance groups regardless of student sample size. Overall, about two thirds of the teachers were stable, while the remaining third (32%) changed. About 45% of the teachers were consistently classified as average teachers (N=1543), 23% of them were consistently classified as significantly above (12%) or below the mean (11%).

The results indicate that student sample size had the largest relative effect on the stability of teacher VAEs, when differences in sample size were large. Sample size variation that is likely to occur within a single cohort, 10 students versus 20 students, produced effects (14%) similar to those we observed for statistical controls (16%). Effects of single versus multiple cohort VAEs on teacher performance group designation fell in the middle (20 to 23%). Part of the effect is due to controlling for within teacher variation, another is an increase in student sample size. Interestingly, the largest effects of sample size on reclassification are very similar to the reclassification percentages we observed when we compared teacher performance between adjacent years (30%), some of which might have been due to sample size if student sample size varied within teachers across the four cohorts.

Finally, we were interested in exploring the stability in VAEs across all fourteen conditions we investigated as part of this study: four different ways of controlling in VAMs for differences in students and other factors teachers might have little control over, four single cohort VAEs, one multiple cohort VAE, and five different student sample sizes. We found that two thirds of all teachers (66%) remained in the same performance group, while one third (34%) of teachers changed. Consistently classified as average performing were 34% of the teachers, 23% were consistently above average, and 9% of the teachers were consistently below average, indicating that performance at the top was more stable than on the low end of the value-added distribution.

Discussion

In this study we examined the relative and combined effects of three basic data and model specifications on the stability of teacher VAEs. We varied the statistical controls included in the models, investigated the stability of single cohort VAEs over time, and compared these yearly estimates with a stable VAE estimated from multiple student cohorts. We also investigated the impact of student sample size on VAE precision and performance group designation. In total, we evaluated VAE stability across fourteen conditions. Understanding the effects of data and model specifications on teacher VAEs is important for policy makers in districts and states as they develop and implement accountability systems that specify the inferences that will be made based on teacher value-added scores. To create policy that is just and fair to teachers, it is important to understand if teacher performance as measured by VAEs is to a large degree a function of data and model

specification, that is, estimated performance is different under different models for a large number of teachers, or if VAEs as a measure of teacher performance are fairly stable.

Several interesting findings emerged from our analyses. Overall, data and model specifications did have an impact on the stability of teacher VAEs. About one third of the teachers changed performance groups across all fourteen conditions. Student sample size had the potentially largest effect on teacher reclassification (32%) when differences in sample size were large. VAEs became twice as precise (standard errors were halved) when we compared sample sizes of 50 or more students to samples of 10 students. Effects of student sample size that might be observed within a single cohort were smaller (14%), and similar to reclassification percentages observed when we included different statistical controls in the models (16%).

Yet, we also observed a fair amount of stability in VAEs in this study. About two thirds of all teachers remained in the same performance group across all fourteen conditions. Consistent with other studies, the top of the value-added distribution was more stable than the bottom (Goldhaber & Hansen, 2010; McCaffrey et al., 2009). This might be important information for policy makers. To be sensible, policy might need to reflect the different levels of stability at the top and at the bottom of the value-added distribution. Further, teachers' relative standing was fairly stable over time. VAE correlations between adjacent years were comparatively high (between .62 and .66) and stable. In that, our results are different from results of other studies, which have reported small to medium size correlations (between .2 and .6) between VAEs from adjacent years, and higher percentages of teachers changing performance groups, with 10 up to 30% of teachers showing up in a performance group on the opposite end of the value-added distribution from one year to the next (Goldhaber & Hansen, 2010; McCaffrey et al., 2009). In contrast, in our sample only 1% of teachers ended up in the opposite group.

These findings raise the question of why we observed greater stability in our analyses than reported elsewhere. For one, the greater stability can be explained by how we measured stability. We considered standard errors, not just the VAE values in teacher performance group designation because we created performance groups that were based on statistical significance. Teachers in either the low or high group had VAEs that were significantly different from the mean, which means that those two groups at either end of the value-added distribution contained a much smaller percentage of the overall distribution than the middle group. The number of performance groups and how they are distributed across the VAE distribution has a direct impact on the number of teachers that will be reclassified because there are much fewer teachers at either ends of the distribution than in the center. For the same value-added distribution, performance groups based on quartiles or quintiles will result in a higher percentage of reclassified teachers because there are more groups, even though differences in performance between groups will be less pronounced. Hence, how we report on stability does matter because it may convey different levels of stability. This might be a valuable insight for policy makers who might consider VAE stability, when evaluating the use of VAEs in a particular jurisdiction.

We chose to distinguish between teacher performance based on statistical significance because it is similar to empirical cut score setting. Teachers whose VAEs are significantly different from the mean might be more likely to be different in other aspects of teacher performance, for example, teaching practices. Ultimately, evaluating stability of VAEs will be most informative if it is done based on actual cut scores set by policy.

Thus, is it fair to conclude that the greater VAE stability in our results is mostly a function of how we measured stability? Not entirely. When comparing adjacent years we

observed only 1% of teachers showing up in the performance group on the opposite side of the value-added distribution in the second year, whereas that percentage would have ranged between 2 and 5 percent based on quartiles, depending on which pair of adjacent years was compared. This is still considerably fewer than the 10 to 30% of teachers reported elsewhere. In this context it might also be noted that we purposefully chose the term reclassification over the term misclassification to describe stability because the latter implies that teachers' "true" VAEs could be known.

We discuss three key findings of this study and their implications for policy makers, and explore some of the underlying characteristics in our data that might explain our results. First, we found that student sample size considerably affected VAE precision and, in turn, teacher classification into performance groups, with up to one third of the teachers being reclassified. Policy makers might need to set standards that specify a reasonable degree of precision, as well as what to do if those standards are not met because the student sample size is too small. A quick way to do this is by specifying a minimum student sample size, which several studies have done (Buddin, 2010; Goldhaber & Hansen, 2010, Harris & Sass, 2010). Another approach is to specify the size of standard errors, a common strategy in psychometrics used for scale construction. Because the quality of a measure is largely determined by how well it can distinguish among members of its target population, the desired precision of scores can be obtained by specifying the size of the standard error. The level of precision in turn is determined by how the scores will be used. High stakes decisions require higher precision and smaller standard errors than low-stakes decisions. In psychometrics, for example, standard errors of size .31 correspond to highly reliable scores (.90) in conventional terms (Embretson & Reise, 2000, p. 270). McCaffrey and colleagues' study is very instructive in this regard. They observed that standard errors decreased considerably with 20 or more students and concluded that 15 or more students should provide reasonably precise VAEs. Our standard errors were of similar size and might provide further support to require 15 or more students per teacher. The average standard error based on 10 students was .13 and for 20 students it was .10. Setting minimum standards for VAE precision will limit the number of teachers for whom VAEs can be estimated. Clearly, it is not reasonable to make the same inferences about individual teachers' VAEs if those VAEs were based on scores from 10 versus 50 students. More work in this area is needed.

Second, in our models we were able to control for more than two thirds (between 68 and 72%) of the overall variance in student scores, with little differences between the four models. We controlled for about two thirds of the variation that was associated with differences among students (62% of the total variance in scores) and some of the variance between classrooms. Whether this indicates that we were able to account for differences in students and their prior learning sufficiently in our models is difficult to say because studies rarely report these values. A key premise in VAMs is that they need to take into account that teachers have different kinds of students whose learning outcomes are likely to vary regardless of the teacher they had. If VAMs are not able to sufficiently control for those differences, VAEs might be biased. If, on the other hand, there is a meaningful relationship underlying student assignment to teachers, controlling for differences based on student characteristics might create bias as well. Investigating the source of the remaining unexplained between-student variance will be an important area of future research. We also noted that student level aggregates did not control for school effects in our data and that those effects were confounded with teacher VAEs. Policy makers might need to consider

the size of school effects in a given jurisdiction and whether they can be controlled for in the statistical models to gauge their impact on teacher VAEs.

Interestingly, we found that how we controlled for differences between students in our models made little difference. Controlling with test scores from the previous year only (mathematics and ELA) was comparable to controlling with two prior years of test scores and combinations between test scores and student background characteristics. Our results are somewhat different from what others have recommended. For example, EVAAS models routinely include test scores from several prior grades for each student to obtain sufficiently reliable estimates of students' prior learning (Ballou, Sanders, & Wright, 2004). One possible reason that might explain our results might be that individual student learning was very stable over time. Correlations between students 3rd, 4th, and 5th grades scores were high, around .8, which suggests that student learning as measured by the standardized test in our sample was very stable and that each test score is essentially a reliable measure of student prior learning. Similarly, correlations between test scores and student background characteristics were of considerable size, replicating much of the information already provided by student scores. It is important for policy makers to recognize that this might not be the case in all districts or states, and that scores from several prior grades could be warranted to account for differences in student prior learning if student learning is quite variable or measured unreliably, but it also shows that not in all contexts scores from multiple prior grades might be required.

In value-added approaches we model student test scores. Because patterns and relationships among the tests scores are likely to affect VAEs, exploring these relationships can provide useful information and guide data and model specifications. Considering the high stability in individual student learning, in this study, we might have been able to include many of the roughly 46,000 students we excluded from our analyses because we did not have their third grade scores. The additional students would have increased the accuracy and precision of many teacher VAEs. Such information might help maximize the number of students used in estimating VAEs, especially in urban school districts such as the one in our study, where student mobility might be greater and obtaining complete score histories for multiple prior grades might be unrealistic for many students. If learning is not measured reliably by the standardized test (i.e., student scores correlate at best moderately over time), perhaps due to the quality of the test itself or considerable within-student variability, then adding scores from multiple prior years might be required to obtain a reasonably reliable estimate of students' prior learning.

Third, single cohort VAEs as a measure of yearly performance were quite stable for adjacent years (r= .62 and r= .66) and still moderately stable over the four years under study (r = .52). The correlations we observed were higher than those reported in other studies (Goldhaber & Hansen, 2010; McCaffrey, et al., 2009), indicating a reasonable degree of stability within teachers over time. In many ways, the stability in VAEs reflects the stability in measured student learning across the four cohorts. Correlations between student test scores from different cohorts were high, between .71 and .75 for 4th grade, and around .80 for 5th grade. Apparently, most teachers had similar kinds of students over the four years of the study, both in terms of incoming mathematics knowledge (grade 4 scores) and at the end of 5th grade. Relating variability in teacher performance over time to variability in incoming students might be one way to better understand the possible sources of within-teacher fluctuations. Do we observe the greatest changes in teacher performance when teachers teach quite different students each year, or is teacher performance just as variable when teachers teach the same kinds of students year after year. Overall, reported changes in

teacher performance seem to be comparable to year-to-year fluctuation reported for other fields where performance can be measured more easily (Goldhaber & Hansen, 2010; McCaffrey et al., 2009).

Although we observed greater VAE stability over time than many studies, correlations of .62 to .66, if interpreted as a measure of consistency over time, are lower than the benchmark value of .8 commonly used in test-retest reliability studies. There might be several reasons why expecting VAE correlations of .8 over time might not be realistic and why using this benchmark value to evaluate stability might not be helpful. If we expect that teacher performance fluctuates somewhat from year to year and changes, even if slowly, over time, we might not expect correlations of a size that would indicate that scores are consistent over time. A key premise of test-retest reliability studies is that either none of the participants changes with regard to the measured variable of interest or that all change at the same rate between the two measurement occasions, and hence that rank-ordering should be preserved. That is also why test-retest reliability studies limit time intervals between measurements to a few weeks or months, certainly not beyond an entire year or multiple years. Assuming that teachers don't change or change at the same rate over the course of a year or longer might not be a reasonable assumption.

In fact, we found evidence of differential changes in teacher performance over time when we considered teacher experience. On average, beginning teachers' performance statistically significantly improved over the four years of study, while performance of teachers with six or more years of experience remained fairly stable. Our results are consistent with findings from studies that examine teacher learning, which have found that improvements in teaching are most notable in the first few years in the classroom (Henry, Fortner, & Bastian, 2012). If policy seeks to address stable, average performance, one implication of our findings might be that VAEs are best used only after a teacher has taught for five years.

In this study we explored the difference between stable, average performance as measured by multiple cohort VAEs and yearly performance estimated from single student cohorts. From a statistical perspective, stable VAEs obtained from multiple student cohorts might be better suited to make inferences about individual teacher performance than single cohort VAEs do simply because they are based on more information. Estimates are more precise because more student data are included and they are more robust because performance variation within teachers has been controlled for. From a policy perspective, however, the choice of single versus multiple cohort VAEs needs to reflect what teacher evaluation and accountability systems seek to reward, which in turn might imply the type of reward. If accountability systems seek to identify who are on average the higher and lower performing teachers in a district or state, multiple cohort VAEs might be preferable and teachers who on average perform above a specified cut score might be rewarded with permanent pay raises. If, on the other hand, such systems seek to identify the higher and lower performing teachers in any given year, single cohort estimates are more appropriate, and bonuses as opposed to pay raises might be a better way to reward yearly performance because it is more variable. Based on our findings, in such a system about half of the teachers who would have received a bonus in one year would not have received a bonus in the subsequent year.

Our investigation of the stability of VAEs revealed both data and model effects but also a fair amount of stability; two thirds of the fifth grade mathematics teachers were stable across all conditions. We found a high level of stability in individual student learning as measured by the standardized test, and a similar level of stability in test scores across

different student cohorts, both of which are likely to have contributed to the stability in VAEs we observed. The higher level of stability in value-added scores in our study might persuade some to conclude that the VAEs we estimated are of sufficient quality to be used in teacher evaluations or accountability systems as measure of teacher performance. Stability, however, doesn't necessarily mean that VAEs are good measures of teacher performance. One alternative interpretation of our findings could be that we are not observing teacher effects but test effects. Is student learning really stable or is this stability a function of the test? More work is needed to understand whether VAEs measure meaningful differences in teacher performance.

References

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

Ballou, D. (2005).Value-added assessment: Lessons from Tennessee. In *Value added models in education: Theory and applications*, edited by Robert Lissetz, pp. 272-303.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Briggs, D. & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/due-diligence.

Buddin, R. (2010). How effective are Los Angeles elementary teachers and schools?, *MPRA Paper* 27366, University Library of Munich, Germany.

Buddin, R. (2011). Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools. MPRA Paper 31963, University Library of Munich, Germany.

Embretson, S. E., & Reise, S. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum Publishers.

Ferrão, M. (2012). On the stability of value added indicators. *Quality & Quantity, 46*(2), 627-637. doi:10.1007/s11135-010-9417-6.

Goldhaber, D. & Hansen, M. (2010). "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." CEDR Working Paper 2010-3. University of Washington, Seattle, WA. Web. Retrieved from http://www.cedr.us/publications.html.

Harris, D. N. (2011). Value-added measures and the future of educational accountability. *Science, 333*(6044),826-827.doi: 10.1126/science.1193793.

Harris, D. N. & Sass, T. R. (2010). What makes for a good teacher and who can tell? Unpublished paper.

Harris, D. N., Sass, T. R., & Semykina, A. (2012). Value-Added Models and the Measurement of Teacher Productivity. Unpublished paper.

Hattie, J. (2003). Teachers make a difference. Paper presented at the 2003 Australian Council for Educational Research Conference. Melbourne, Australia. Retrieved November 11, 2012, from

http://www.education.auckland.ac.nz/webdav/site/education/shared/hattie/docs/teachers-make-a-difference-ACER-(2003).pdf

Henry, G. T., Fortner, C. K., & Bastian, K. C. (2012). The effects of experience and attrition for novice high-school science and mathematics teachers. *Science, 335*(6072), 1118-1121. doi:10.1126/science.1215343.

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

Hiebert, J. & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 371 - 404). Charlotte, NC: Information Age Publishing.

Kane, T., Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper 14607.

Kane, Thomas J. and Douglas O. Staiger. (2002a). "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." In Brookings Papers on Education Policy,2002, edited by D. Ravitch. Washington, DC: Brookings Institution.

Kane, T. J., & Staiger, D. O. (2002b). The promise and pitfalls of using imprecise school accountability measures. Journal of Economic Perspectives, 16, 91-114.

Koedel, C., & Betts, J., 2007. Re-Examining the Role of Teacher Quality in the Educational Production Function, Working paper, University of California, San Diego.

McCaffrey, D., Sass, T., Lockwood, J., and Mihaly, K. (2009). The Inter-Temporal Variability of Teacher Effects Estimates, Education Finance and Policy, Fall 2009, Vol. 4, No. 4: 572–606.

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

Measures of Effective Teaching Project (2010). Learning about teaching: initial findings from the measures of effective teaching project. MET Project Research Paper. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

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

National Council on Teacher Quality. (2011). State of the states. Trends and Early Lessons on Teacher Evaluations and Effectiveness Policies. New York. Retrieved 09/26/2012 from http://www.nctq.org/p/publications/docs/nctq_stateOfTheStates.pdf.

NCLB (2001). No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425.Noell, G. H., & Burns, J. L. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*, 37-50.

Papay, J. P. (2011). Different Tests, Different Answers. *American Educational Research Journal, 48*(1):163-193.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29 (1), 121–129.

Roderick, M., B. A. Jacob, and A. S. Bryk. 2002. The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation and Policy Analysis* 24(4)(Winter):333–357.

Rothstein, J. (2007). *Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference*,Working Papers 1036, Princeton University, Department of Economics, Center for Economic Policy Studies.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics,* 125(1), 175-214.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary school. *Teachers College Record, 104*(8), 1525-1567.

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

Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee Value-Added Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In Millman, J. (ed.), Grading Teachers, Grading Schools, Thousand Oaks, CA: Corwin Press.

Schochet, Peter Z. and Hanley S. Chiang (2010). Error Rates in Measuring Teacher and School
Performance Based on Student Test Score Gains (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

The New Teacher Project. (2007). *Teacher hiring, transfer and assignment in Chicago Public Schools*. New York: Author. Retrieved January 25, 2008, from http://tntp.org/assets/documents/TNTPAnalysis-Chicago.pdf?files/TNTPAnalysis-Chicago.pdf

About the Authors

Nicole B. Kersting
University of Arizona
Email: nickik@email.arizona.edu
Nicole B. Kersting is an Assistant Professor in the Department of Teaching, Learning and Socio-cultural Studies and a faculty member of the Interdisciplinary Graduate Program in Statistics at the University of Arizona. Her research is focused on the measurement of different aspects of teacher
quality: Teacher knowledge, instructional quality, and student learning, and value-added approaches to estimate teacher performance. She developed and validated a novel approach to measure teacher knowledge in mathematics that is based on teachers' analyses of classroom video clips.

Mei-Kuang Chen
University of Arizona
Email: kuang@email.arizona.edu
Mei-kuang Chen is a postdoctoral researcher in the College of Education, University of Arizona. She graduated from the Department of psychology, with a focus on Program Evaluation and Research Methodology. Her research interest is in applying critical thinking, research methods, and statistical tools in various areas to gain a better understanding of human phenomena.

James W. Stigler
University of California Los Angeles
Email: Stigler@ucla.edu
James W. Stigler is Professor of Psychology at UCLA. His research focuses on understanding processes of teaching and learning, especially of mathematics and science, from kindergarten through college. James was Director of the TIMSS video studies and CEO of Lessonlab. Stigler is best known for his observational studies of mathematics and science teaching, and has pioneered the use of multimedia technology for the study of classroom instruction.

Acknowledgements

We would like to thank Kilchan Choi for his support in the early stages of this project. His experience in value-added modeling approaches helped us understand the technical challenges involved in these models. We also would like to thank Lee Seechrest with whom we have had many interesting discussions throughout this work and who has given us valuable feedback in the writing of this article. All mistakes or errors are responsibility of the authors. This work is supported by a grant from the National Science Foundation (NSF award # 0949241).

About the Guest Editor and Assistant Guest Editors

Guest Editor
Dr. Audrey Amrein-Beardsley

Arizona State University
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.

Assistant Guest Editor
Dr. Clarin Collins
Virginia G. Piper Charitable Trust
clarin.collins@asu.edu
Clarin Collins recently completed her Ph.D. in Educational Policy and Evaluation from Arizona State University, with an emphasis in research methods. Via her dissertation, she examined teachers' understandings of and experiences with the SAS Education Value-Added Assessment System (EVAAS) in the Houston Independent School District where it is used to evaluate teacher effectiveness. Clarin is currently a Research and Evaluation Officer at the Virginia G. Piper Charitable Trust in Phoenix.

Assistant Guest Editor
Dr. Sarah Polasky
Arizona State University
sarah.polasky@asu.edu
Dr. Sarah Polasky is the Value-Added Specialist for the Arizona Ready-for-Rigor Project, a Teacher Incentive Fund Grant, within the Mary Lou Fulton Teachers College. Her current research interests include the development and implementation of value-added measurements and systems using high-stakes test data, assessment in early childhood education, the use of alternative achievement (e.g., district benchmarks, formative assessments) and non-achievement (i.e., developmental) data for value-added analysis, as well as the impact of socioemotional and neurological development of young children on their short- and long-term academic achievement.

Assistant Guest Editor
Edward F. Sloat
Mary Lou Fulton Teachers College, Arizona State University; Dysart Unified School District, Surprise, Arizona
esloat@asu.edu
Mr. Sloat is currently employed as the Director of Research and Accountability at Dysart Unified School District located in Surprise, Arizona and a doctoral student in the Leadership and Innovation Program within the Mary Lou Fulton Teachers College, Arizona State University. Mr. Sloat has served as Deputy Associate Superintendent for Research and Evaluation within the Arizona Department of Education, the Director of Research, Planning, and Assessment for the Peoria (Arizona) Unified School District, and as Director of Research and Assessment at the Glendale (Arizona) Elementary School District. He regularly contributes to state technical and policy working/advisory groups concerning assessment design and accountability systems and is past President of

the Arizona Education Research Organization. Mr. Sloat holds a Master's Degree in Applied Economics from the University of Arizona, concentrating in econometric methods and management information systems. His academic interests focus on value-added modeling, education accountability and evaluation systems, data-driven instructional planning, applications of measurement theory, and research methods.

SPECIAL ISSUE
Value-Added: What America's Policymakers Need to Know and Understand

education policy analysis archives

Volume 21 Number 7 January 29, 2013 ISSN 1068-2341

education policy analysis archives
editorial board

arquivos analíticos de políticas educativas
conselho editorial