Educ Asse Eval Acc (2014) 26:203–222
DOI 10.1007/s11092-013-9183-9

A comparison between value-added school estimates and currently used metrics of school accountability in California

Loris P. Fagioli

Received: 27 May 2013 / Accepted: 30 December 2013 /
Published online: 15 January 2014

Abstract This study compared a value-added approach to school accountability to the currently used metrics of accountability in California of Adequate Yearly Progress (AYP) and Academic Performance Index (API). Five-year student panel data (N= 53,733) from 29 elementary schools in a large California school district were used to address the research questions. Results show the strong relationship between AYP and API to student background measures. Schools with a majority of students from low socioeconomic background lagged far below schools from more affluent context. Results from the value-added approach however, showed a strongly diminished relationship to student background. Under this model, several schools from a low socioeconomic context can be seen as high achieving. Additionally, little evidence was found that high levels of student achievement negatively affect school value-added scores. Schools that enroll large proportions of advanced students, which often do not show positive growth across years are not penalized under a value-added approach. Recommendations for policy and future research are discussed.

Keywords Value-added models · School accountability · Academic performance · Adequate Yearly Progress (AYP) · Academic Performance Index (API)

The introduction of the No Child Left Behind (NCLB) Act in 2001 (NCLB 2001) increased the focus on accountability in education in the USA. NCLB requires states to develop accountability systems in mathematics and reading and monitor student achievement toward proficiency in state-developed standards in those areas. The emphasis on accountability and its “high stakes” measures of student achievement puts schools and their staff under considerable pressure to perform at high levels. However, some critics believe that the measures used for accountability are not real measures of success and are biased towards already high performing schools (Lee 2003; Linn 2006, 2008; Manwaring 2010). Recently, the US federal government announced in 2009 that

L. P. Fagioli (✉)
Claremont Graduate University, Claremont, CA, USA
e-mail: loris.fagioli@cgu.edu

states can apply for $4.35 billion Race to the Top funds for education. In order to qualify, states have to meet certain criteria to be eligible for consideration. As part of the criteria, states are required to use a new evaluation system. This methodology called value-added modeling tries to attribute changes in students' achievement to their school or classroom taking into account several factors that vary between students and schools (Chudowsky et al. 2010). Value-added scores can be used to evaluate schools and teachers and whether their average student performance exceeds or falls behind a similar classroom or school. Proponents of this methodology believe that value-added models (VAMs) are a fairer way to assess performance compared with the more traditional policies that only look at a "snapshot" of current student achievement which is often related to student background (Chudowsky et al. 2010; Harris 2011).

The purpose of this paper is thus to compare the new methodology of value-added modeling to the traditional measures of accountability. The focus is, specifically, on how California measures of school accountability relate to value-added measures of schools in one California school district and which measure is most highly related to student background such as socioeconomic status (SES), ethnicity, and language proficiency.

1 Accountability in California

California's poor performance in the early 1990s on the National Assessment of Educational Progress made improving student performance a priority for legislators (Hart and Brownell 2001). By 1998, core academic standards were defined in English Language Arts (ELA), Mathematics, Historical–Social Science, and Science. These standards defined what students were expected to know at each grade level. The California assessment system uses these standards as the criterion reference for its tests. The Academic Performance Index (API) is part of this assessment system and one of the most important measures of the Public School Accountability Act of 1999. This school measure ranges from 200 to 1,000 and is a weighted average of student scores across the content areas (for more information on the calculation of API scores, see California Department of Education 2013a, b). The performance target for all schools is a score of 800 (California Department of Education 2013c). Schools that do not meet the required level get an annual target API which is 5 % of the difference between the current school API and the target score of 800 (Mintrop and Trujillo 2007). In addition, several relevant student subgroups such as socioeconomically disadvantaged students, English learners, students with disabilities, and several ethnic groups need to achieve the target score (California Department of Education 2013c).

US Policy has mirrored California efforts toward greater accountability. The goal of the federally mandated NCLB Act of 2001 is that all students reach levels of proficiency by the 2013–2014 academic year (NCLB 2001). The act requires states to report and monitor the progress towards that goal and set a timeline for Adequate Yearly Progress (AYP) for all schools. The AYP consists of annual measurable objectives (AMOs) which in part have state specific definitions and vary from state to state (Erpenbach and Forte 2008; Kim and Sunderman 2005). In California, the four requirements for the AYP are: 95 % student participation rate overall, a defined percentage of students scoring at the proficient or above level in ELA and mathematics, an API level of 740 or 1 point growth, and meeting graduation rate targets for high schools (California Department of Education 2013d). A school fails AYP if a single subgroup requirement or one AMO is not met. Schools that do not meet AYP criteria in two consecutive years are defined as in "need of improvement." Schools that receive Title I funding (a federal program supporting schools with a high percentage of students from economically disadvantaged families) can be put on program improvement, which entails several sanctions depending on how many years a school is in the program. Sanctions can range from students having the option to transfer (1st year of improvement) to restructuring, conversion to a charter school, or hiring of all new staff in their 5th year of improvement (Kim and Sunderman 2005).

Both API and AYP are therefore important diagnostic tools in California to determine school performance. However, high-poverty schools face tremendous challenges in meeting these goals (Kim and Sunderman 2005). Evaluating schools based solely on their student's mean achievement is fraught with problems. Mainly, it fails to account for student background and is subject to selection bias. In other words, schools that serve students from poor and minority backgrounds are disadvantaged from the start. Although some schools might achieve large student gains, their average API score is likely to be lower and these schools might be at a higher risk of being labeled "failing" compared with schools in more affluent areas. Several studies have shown that mean differences between schools often stem in large parts from differences in student background (Kim and Sunderman 2005; Linn 2000; Raudenbush 2004). In recent years, an alternative evaluation system has become more prominent that isolates the contribution of a school or teacher to student learning. The following sections describe this methodology in more detail.

2 Value-added modeling

Research and educators seem to struggle to come up with definitions of a good education (Goldhaber and Brewer 2000; Goldhaber 2002; Wilson and Floden 2002). Over the past two decades, a line of research has emerged that attributes value to the ability to improve student achievement. This value-added approach differs from the more traditional mean or 'raw' score approach in several ways. In its most basic form, a value-added score is just the difference or *gain* score of a student from 1 year to the next. However, this simple approach is questionable in several respects (Chudowsky et al. 2010) and has evolved to more complex models now called VAMs. Just comparing gain scores among schools is problematic, as schools vary in terms of several factors not under their control and a comparison between unequal schools would be unfair. The alternative is therefore to compare a school to other schools that have similar conditions. However, instead of trying to match schools in terms of their similarity (which could be difficult or even impossible), the VAMs attempt to predict an expected achievement level for each school but account or adjust for several known factors that have an impact on achievement but are outside of a school's control (such as school resources and student background factors). Adding controls such as SES, ethnicity, or language learner status makes gain scores more comparable and helps in not penalizing or favoring certain students and schools. Based on this predicted achievement and controlling for several factors, a school's value-added score can be calculated. If a school performs above expectation, the value-added score for that school would be higher than for schools who perform below expectation (see Harris 2011).

VAMs are flexible in terms of controlling for several differences that might affect student achievement. However, quantifying the ‘added value’ is quite intricate. The professional literature is full with debates of the best ways of ‘disentangling’ the true value-added effect from other factors that have an influence on student scores and how choosing a specific model affects these estimates (Chudowsky et al. 2010; Guarino et al. 2012; McCaffrey et al. 2003, 2004; Newton et al. 2010).

Researchers have yet to agree which model or approach leads to the best estimate or if estimating some value based on student scores is possible at all. Much of the literature on the topic is highly technical and difficult to understand for nonexperts. It is thus not surprising that VAMs have been branded by some as “mathematical intimidation” (Ewing 2011). In addition, some publications point to VAMs containing a considerable amount of error (e.g., Guarino et al. 2012; Schochet and Chiang 2010) and having major flaws in several other respects specifically concerning the nonrandom assignment of teachers to classrooms (Braun 2005; Briggs and Domingue 2011; Rothstein 2008). Large-scale independent studies tend to support VAMs while at the same time pointing to some of its concerns and limitations (Bill and Melinda Gates Foundation 2010; Chudowsky et al. 2010; McCaffrey et al. 2003). Despite the criticism that these models are biased and unfair evaluations and do not capture the true value of schools or teachers, it is maybe ironic that precisely this methodology contributed to findings showing that teachers are the most important school factor in raising student achievement (Rivkin et al. 2005; Rowan et al. 2002; Sanders et al. 1997). Other often-cited results include simulations that show how consecutive high quality teachers could erase the achievement gap (Rivkin et al. 2005; Sanders and Rivers 1996) or how getting rid of the bottom 5–10 % of teachers could erase the US handicap in international comparisons (Hanushek 2009).

Value-added scores can be computed at several levels. Most research in the US is focused on teacher-level analyses. However, a recent report by the US Department of Education showed that teacher-level value-added scores contain a considerable amount of error (Schochet and Chiang 2010). Since student test scores are highly related to background factors, several years of data are needed to estimate the teacher effect on these scores. Even with 3 years of data, the study showed that there is a 26 % chance of misclassification of a teacher. That is, about one in four teachers could be classified as underperforming when in reality they show average performance and one in four teachers could be classified as average when they actually performed above or below the mean. Another simulation study shows that misclassification rates can vary widely depending on model and type of student–teacher assignment (Guarino et al. 2012). Schochet and Chiang (2010) showed that the rate of misclassification can be reduced if the value-added score is computed at the school level, thereby increasing the sample size and reliability of the estimates. Indeed, error rates at the school level are 5–10 % lower than teacher-level scores (Schochet and Chiang 2010). Given the more accurate classification of school-level scores, the authors suggest using a value-added approach for school accountability policies. However, little research has been published on the topic of school value-added in US context (for international evidence see below). This paper is therefore an attempt at comparing such an approach with the currently used school accountability metrics in California.

School and teacher accountability is a hotly debated topic in education. Maybe the best known public example in California of the use and controversy surrounding the methodology of VAMs are the articles and ongoing analysis from the *LA Times* that publishes the value-added scores for teachers in the LA Unified School District (LAUSD) (LA Times 2011). But despite some fierce criticism, many leaders such as the US Secretary of Education, Arne Duncan (Cody 2012), or LAUSD superintendent John E. Deasy (Blume 2013) favor accountability measures that are based on VAMs. In addition, several states and school districts have already implemented or are planning implementation of value-added measures for schools or teachers such as Florida, Tennessee, North Carolina, New York City public schools, and D.C. public schools (Butrymowicz and Garland 2012; Goe 2008). However, there is still considerable debate over the reliability and validity of VAMs as well as which model gives the most accurate scores (Braun 2005; Harris 2011; McCaffrey et al. 2004; Sanders et al. 2009). These questions are vital, as the choice of model affects the estimate and essentially determines if a teacher or school is labeled effective or not. Nevertheless, proponents of VAMs stress that despite its flaws, VAMs still support a more just and fair way of evaluation compared with the current model of just looking at mean scores or “snapshots” of student achievement at one point in time (Harris 2011).

VAMs are in use or under consideration in several countries but feature most prominently in the UK and the USA (OECD 2008). In the UK for example, value-added measures for schools were piloted in the 1990s (Saunders 1999). The models went through several versions and schools are now evaluated under a Contextualized Value-Added model which incorporates a range of background factors in its calculation (Kelly and Downey 2010; Leckie and Goldstein 2009; Ray et al. 2009). The results are published annually in performance tables to assess each school’s effectiveness, hold schools publicly accountable, and can be used as a help in school choice decisions (Leckie and Goldstein 2009). Research on value-added school measures is also available from countries such as Australia (Downes and Vindurampulle 2007), the Netherlands (Timmermans et al. 2011), New Zealand (Strathdee and Boustead 2005), and Poland (Jakubowski 2008). A report by the OECD also discusses the potential of implementation of value-added school assessments in several other OECD countries (OECD 2008).

3 Research questions

The purpose of this study is to use a value-added methodology to evaluate school performance and compare it to the currently used metrics of school accountability. The following research questions will be addressed specifically:

1. What is the relationship between value-added scores, AYP, and API of schools in an urban California school district?
2. Comparing school value-added scores, AYP, and API, which school measure is most highly related to student background?
3. Do large numbers of high-achieving students negatively affect a school’s value-added score?

Based on the theory of value-added and previous research on student achievement measures, we would expect a robust relationship between student background and AYP and API. This relationship is expected to be much reduced for value-added measures as student background is part of the control mechanisms built into the methodology. Nevertheless, it is unclear how a value-added accountability could affect schools in areas of higher affluence, which might not be able to raise student scores substantially because of the already high levels of achievement.

4 Methods

4.1 Data

The dataset used for this study stems from a large, predominantly Hispanic school district in Southern California with a yearly enrollment of about 30,000 students (Pre-Kindergarten to 12th grade). The available dataset contains panel data for 5 years (2006/2007–2010/2011) for grades two to six for all elementary schools in the district (*N*=29). Data were available on student achievement (ELA and Mathematics), student background (race, English language learner status[1], disability status, and eligibility for free or reduced lunch[2]).

The dataset contained information on 56,806 students. The following cases were excluded from this sample: students with missing values on ELA or Math scores (2 %), students from classrooms with fewer than 15 or more than 50 children (probable misclassification in dataset or specialized classrooms, 2 %) and students from special education classrooms (classrooms with only students with a diagnosed disability, 0.6 %). Students in special education classrooms receive specialized or supplemental instructions (e.g., additional teacher aids, one-on-one instruction) and often take the California Modified Testing, which is different from the California Standards Test (CST). Therefore, the sample was limited to students not in special education[3] leaving a final sample of *N*=53,733 students in 29 schools over 5 years.

4.2 Variables

The dependent variables are the CST Math and ELA achievement scores. Student background variables include binary indicators of a student's race (Asian/White, Hispanic, or other race[4]), if student is from low-SES (receives free or reduced lunch), if student has diagnosed disability (SWD; student has 504 plan or an Individualized Education Plan), and if a student is classified as an English Learner (EL). Unfortunately, information on other important student characteristics such as age, gender, parental education, and other school and staff characteristics (e.g., teacher

[1] Students who are not proficient in English are classified as English language learners.

[2] Free or reduced lunch is a federally assisted school lunch program for children from households with low incomes.

[3] Harris (2011) cautions against an approach that does not include Special Education as part of an accountability system. However, this paper does not attempt to evaluate specific schools or suggest a specific VAM for policy implementation, but is interested in the more broad relationship between student background and school accountability. More research is certainly needed to evaluate the effect of the inclusion or exclusion of Special Education in a value-added approach to school accountability.

[4] The three categories were defined mainly because of the sample sizes of each group in the district. The district is predominantly Hispanic and other races and ethnicities were collapsed for meaningful interpretation. White and Asian ethnicities were collapsed into one category because of limited size but similar achievement levels, and all remaining ethnicities were defined as other.

experience and teacher qualifications) were not available in the dataset. The inability to include these variables is a limitation of this study (see discussion below).

4.3 Model

This study uses the following model to estimate school quality:

$$T_{jt} = T_{jt\text{-}1}\beta_1 + \boldsymbol{X}_{jt}\beta_2 + \boldsymbol{C}_{jt}\beta_3 + \phi_{\boldsymbol{j}} + \varepsilon_{jt}$$

Where T represents a student's test score for school j in year t; $\boldsymbol{X}$ is a vector of student characteristics, $\boldsymbol{C}$ is a vector of classroom characteristics; ϕ are school fixed effects; β are estimated effects; and ε are unobserved characteristics that are related to the outcome.

The student characteristics include student race, low-SES, SWD, and EL. Classroom level variables were included to control for peer effects: if classroom was multigrade (several grades taught simultaneously in one classroom), average previous classroom achievement in Math or ELA, percent of students from low-SES, racial composition of classroom, percent of EL students, percent of SWD, classroom size, and school size.

This study uses a dynamic OLS approach to estimate the value-added scores. In several simulations with different nonrandom teacher to student assignment models, the dynamic OLS approach was the most stable with the lowest misclassification rates of all models under consideration (Guarino et al. 2012). This model also uses previous achievement as a control instead of using gain scores as the dependent variable. Guarino et al. (2012) showed that lagged models provide consistent effects even though some researchers have voiced concern with that approach (see Rothstein 2007). Similarly, models that include multilevel fixed effects (e.g., student and school fixed effects) have been shown to have higher year-to-year volatility (Goldhaber et al. 2012) and therefore only school fixed effects are included. The model also does not employ empirical Bayes shrinkage as some research points to comparable results between OLS and empirical Bayes estimations (Schochet and Chiang 2010).

The school value-added scores were estimated in two steps. First, the model was estimated separately by subject (Math and ELA) and year. Second, the individual school estimates by subject were averaged by year to obtain a single value-added score for each school. With the available 5 years of data, the model estimated four value-added scores per school (see Appendix for regression results).

5 Results

5.1 Descriptives

Table 1 shows the descriptives (5-year averages) of the student, classroom, and school characteristics[5]. The district showed large variation on several dimensions. Student achievement scores ranged from the lowest possible score (150) to the highest possible scores (600) but also showed large variation at the classroom and school level. Even

[5] Student characteristics were averaged at the classroom and school level. Average school-level characteristics are reported for reference and are not included in the model.

Table 1 Descriptives of student, classroom, and school characteristics

Variable	Number[a]	Mean	SD	Min	Max
Math achievement (CST)	53,733	359.39	77.88	150	600
ELA achievement (CST)	53,733	340.17	55.17	150	600
Asian/White	53,733	0.120	0.325	0	1
Hispanic	53,733	0.817	0.387	0	1
Other race	53,733	0.063	0.243	0	1
Low-SES	53,733	0.716	0.451	0	1
SWD	53,733	0.046	0.210	0	1
English Learner	53,733	0.418	0.493	0	1
Average Math achievement in classroom	2,021	359.335	44.503	212.917	555.839
Average ELA achievement in classroom	2,021	339.921	30.056	241.833	510.516
Class Size	2,021	26.956	4.979	15	35
% low-SES in classroom	2,021	0.717	0.247	0	1
% SWD in classroom	2,021	0.046	0.058	0	0.355
% Asian/White in classroom	2,021	0.118	0.182	0	0.968
% Hispanic in classroom	2,021	0.819	0.206	0	1
% other race in classroom	2,021	0.063	0.069	0	0.42
% EL in classroom	2,021	0.427	0.246	0	1
Multigrade classroom	2,021	0.099	0.299	0	1
School API	144	761.083	75.106	656	969
School API growth[b]	116	8.368	25.166	−63.815	126.703
AYP	144	0.403	0.492	0	1
School size	144	394.486	126.956	177	719
% low-SES in school	144	0.691	0.246	0.091	0.944
% SWD in school	144	0.066	0.050	0	0.170
% Asian/White in school	144	0.136	0.188	0.011	0.810
% Hispanic in school	144	0.794	0.207	0.157	0.979
% Other race in school	144	0.070	0.054	0.008	0.229
% EL in school	144	0.403	0.207	0.019	0.8
% Multigrade classroom in school	144	0.112	0.085	0	0.364

Five-year averages reported

CST California Standards Test, *ELA* English Language Arts, *Other race* not Hispanic, Asian, or White, *Low-SES* students receiving free or reduced lunch, *SWD* students with disability

[a] Represents 5-year total sample size except for school API growth (4-year total sample size)

[b] Sample size smaller from API Growth since previous year API only available for 4 years

though the district was predominantly Hispanic (82 %) racial school composition was not evenly distributed. Schools ranged from 16 to 98 % Hispanic and from 1 to 81 % Asian or White students. Similarly, most students were low-SES (72 %), but some classrooms had no students from low-SES and schools ranged from 9 to 94 % low-SES.

Overall about 5 % of students were SWD and about 42 % were EL. But again, there was variation in how these students were distributed across classrooms and schools. Class size within schools ranged from 15 to 35 students and school size also varied from a student body of 177 to 719 students. School API ranged from 656 to 969. School API growth from the previous year ranged from a decrease of about 64 points to an increase of about 127 points. About 40 % of schools met their AYP goals.

These descriptives point to a large variation within the district between classrooms and schools in terms of racial, socio-economic, and other student background composition. This district with its large variation therefore lends itself for the intended analysis of this paper comparing the impact of VAMs on schools from different contexts.

5.2 Stability of value-added scores across years

One criticism of value-added modeling pertains to the sometimes low correlations or stability of value-added scores from 1 year to the next. Some research found modest correlations for teacher value-added scores ranging from 0.2 to 0.6 (Goldhaber and Hansen 2010, 2013; McCaffrey et al. 2009). Therefore, in order to determine the stability of school value-added scores the between year stability was analyzed as shown in Table 2. The rank ordering of the value-added scores for a school was highly consistent with Spearman's rank correlations ranging from 0.95 to 0.98 in Math and from 0.85 to 0.95 in ELA. Furthermore, the Math and ELA rank scores were highly correlated (r=0.939)[6] as well, indicating a consistent contribution of both subjects to the overall school value-added score.

5.3 Relationship between value-added, AYP, and API

The first research question of this paper focused on the relationship between value-added scores, AYP, and API. Table 3 summarizes the average within-year correlations between the three accountability measures. Additionally, a fourth measure was added that represents the API growth score (API growth from previous year). The value-added scores only showed a significant low correlation of 0.24 with API but the relationships with AYP and API growth were not significant. However, AYP and API were significantly correlated (r=0.62). Additionally, AYP and API growth also showed a significant relationship of r=0.33. The relationship between AYP and API growth could be expected, as the AYP is based in part on a school's API growth.

Based on this analysis, value-added scores do not show any strong relationship between currently used metrics of school accountability. Based on this simplistic analysis, the value-added analysis seems to be measuring a different quality, whereas the AYP and API are measuring a similar quality of schools.

5.4 Relationship of value-added, API, and AYP with student background variables

The second research question focused on the relationship of student background to measures of school quality. Table 4 shows the results between the value-added, AYP, API, and API growth of a school and its relationship to background variables (low-SES, English Learners, and racial composition). API is highly correlated with low-SES and minority composition of the school (r=−0.83 to −0.87). The negative relationship

[6] Represents average rank correlation within-year across 4 years.

Table 2 Stability of rank order of value-added scores between years

	Math value added			ELA value added		
	2nd year	3rd year	4th year	2nd year	3rd year	4th year
3rd year	0.970			0.925		
4th year	0.963	0.976		0.887	0.907	
5th year	0.950	0.968	0.983	0.850	0.854	0.950

All coefficients significant at α=0.05, Spearman's rank correlation coefficient reported

implies that schools with higher proportions of low-SES students and minority students have lower API scores. AYP also showed significant relationships with these variables although of lower magnitude (r=−0.51 to −0.56). The relationship of value-added with these same variables however gives a different picture. The magnitude of the relationship is substantially smaller with correlations ranging from −0.27 to −0.33. API growth showed no significant relationship to these student background variables. Even though the correlations did not reach significance, using a simple school growth measures for accountability could be cause for concern as API growth had a positive relationship in the sample with low-SES, EL, and Hispanic students, but a negative relationship with Asian/White students.

The reduced relationship between student background variables and value-added scores compared with API scores is also presented visually in Figs. 1 and 2 (4-year school averages shown). When looking at Fig. 1, there is a clear distinction between low- and high-SES schools when measured on API. All top performing schools are on the left (few low-SES students) and all those schools performed above the California target of an API of 800. Only one school on the right side (majority low-SES students) performed above that level. All other schools performed below target level but also serve a majority of low-SES students.

When looking at Fig. 2, however, with a value-added approach, this relationship between student background and school achievement is weaker. Both low- and high-SES schools show high performance. While majority high-SES schools still perform average and above, there are several low-SES schools that perform 200 or more points above expectation.

Table 3 Average correlations between value added, API, and AYP

	Value added	AYP	API
AYP	0.181		
API	0.241*	0.623*	
API growth	−0.078	0.334*	0.161

Coefficients represent the average within-year correlations

*p<0.05

Table 4 Correlations of value added, API, and AYP with student background variables

	Value added	API	AYP	API growth
Low-SES	−0.266*	−0.843*	−0.508*	0.125
% English learners	−0.278*	−0.825*	−0.540*	0.097
% Hispanic	−0.334*	−0.860*	−0.556*	0.099
% Asian/White	0.284*	0.867*	0.536*	−0.087

Coefficients represent the average within-year correlations

*p<0.05

The difference between the two approaches can also be seen in Table 5 which lists the top and bottom schools based on their average value-added rank. Looking at the five top schools, two of those were also in the top five based on API ranking. The other three top performing schools however, ranked at the bottom of API ranking. These three schools have a majority of low-SES, EL, and Hispanic students. Looking at the bottom five performing schools, all have majority low-SES, EL, and Hispanic students, as well as lower rankings based on API. These schools seem to truly underperform regardless of accountability method.

5.5 Relationship of high-achieving students with school VA scores

One question unexplored so far is if schools are penalized for their good students (i.e. large growth is not possible with high-achieving students which could result in low

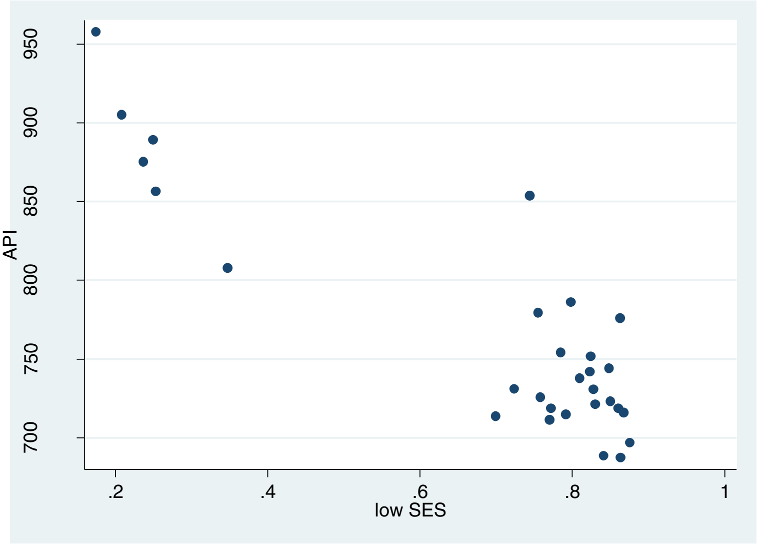

Fig. 1 Scatter plot of 4 year averages between a school's API and percent of students from low SES (N=29 schools)

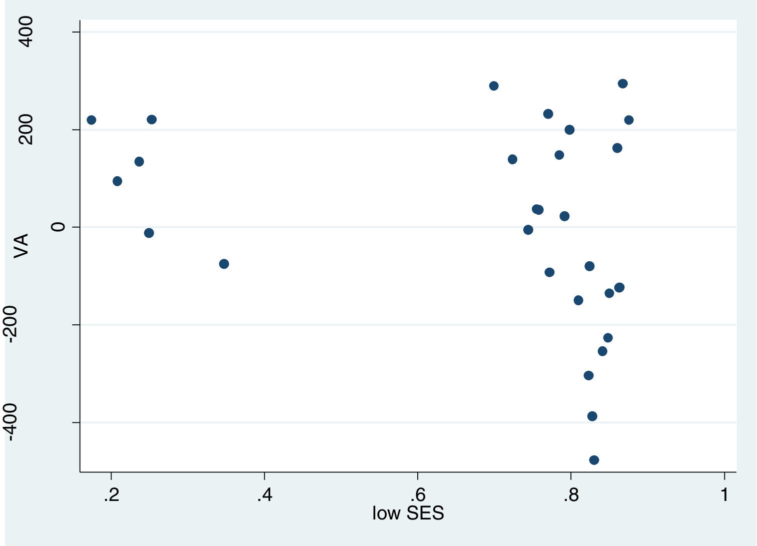

Fig. 2 Scatter plot of 4-year averages between school value-added scores and percent of students from low SES (*N*=29 schools)

value-added scores). To investigate this question, student cohorts were categorized according to their five proficiency levels at the beginning of each school year: far below basic, below basic, basic, proficient, and advanced (California Department of Education 2013e).[7] The average CST growth during the year of these five levels was cross-tabulated with the school value-added score (low, medium, and high) to see the average growth achieved by these five groups based on their incoming level of achievement. In addition to the overall impact, schools were classified as low-SES (50 % and more low-SES students) and high-SES (fewer than 50 % of students from low-SES). As can be seen in Table 6 in the overall section, there are differences between the low, medium, and high value-added schools in terms of their student growth. For instance, low value-added schools on average raised far below basic students by 31.9 points, whereas high value-added schools raised them by 34.8 points. Similarly, top quintile students on average dropped 20.2 points in low value-added schools, whereas in high value-added schools these students only dropped 13 points. The positive growth by nonproficient students, and the negative growth by proficient/advanced students could be expected due to regression to the mean[8] (Barnett et al. 2005). However, the difference between the schools in

[7] Proficiency levels were computed based on the average CST score of Math and ELA.

[8] Every test contains some measurement error (e.g., luck, item guesses, mental state on day of testing, etc.). With test scores at the top and the bottom, this measurement error is likely to change when retested. Top scores tend to decrease when retested and bottom scores tend to increase. This phenomenon is called regression to the mean (for more information, see Stigler 1997)

Table 5 Top and bottom five schools based on average value-added rank

VA rank	API rank	API	Hispanic (%)	Asian/White (%)	Low-SES (%)	EL (%)	VA
1	23	716.0	95.6	2.9	86.7	72.5	293.7
2	25	713.8	70.0 %	9.1	70.0	25.6	289.7
3	26	711.5	86.3	7.5	77.0	45.8	232.3
4	5	856.5	52.7	41.0	25.3	8.5	220.3
5	1	957.8	22.7	72.0	17.4	3.4	219.7
25	13	744.0	93.2	3.3	84.8	62.7	−227.5
26	28	688.3	94.8	1.8	84.1	58.3	−254.3
27	14	741.8	90.3	8.4	82.3	45.4	−304.0
28	17	730.8	95.4	2.9	82.8	52.7	−387.1
29	20	721.0	91.5	3.6	83.0	54.4	−477.7

Values based on 4-year averages

VA Value-Added, *API* Academic Performance Index, *EL* English Learner

terms of their value-added is how this regression to the mean manifests itself. High value-added schools have a slower rate of regression for top performing schools, but a faster regression at the bottom.

Based on these descriptives that show on average negative growth for high-achieving students, one could assume that higher proportions of high-achieving students in a school would reduce the overall student achievement scores and hence negatively affect a school's value-added score. However, as Table 7 shows, this assumption of negative impact is not supported when looking at correlations of school value-added scores and percent of advanced students. There is actually a slight positive relationship between larger numbers of advanced students and school value-added scores ($r=0.256$) and some evidence of a slight negative effect of larger proportions of below basic students ($r=-0.336$).

5.6 Limitations

This study was limited in several respects. Several key variables such as gender, parental education, and teacher qualifications were not available in the data and are hence not incorporated in this model. These controls could add valuable information and therefore increase the precision of the value-added estimates. Additionally, the data stemmed from a predominantly Hispanic community and the majority of schools can be considered low-SES. The results are thus not representative of California as a whole and it is not clear if these results would be replicated with other schools or districts. The data were also limited to elementary schools and excluded any high schools. Future analyses should try to incorporate additional meaningful controls that limit the influence of factors

Table 6 Average student CST growth by incoming CST level, school value added, and school SES

	Incoming CST level	School value-added categories								
		High-SES schools			Low-SES schools			Overall		
		Low	Medium	High	Low	Medium	High	Low	Medium	High
Average CST growth in level	Far below basic	No schools	53.8	40.9	31.9	38.2	34.7	31.9	39.7	34.8
	Below basic		18.9	36.2	13.3	14.5	14.4	13.3	14.9	16.2
	Basic		1.6	11.8	1.9	2.8	0.1	1.9	2.6	2.2
	Proficient		−4.4	1.3	−8.2	−4.7	−9.6	−8.2	−4.7	−5.9
	Advanced		−15.0	−10.2	−20.2	−18.8	−18.0	−20.2	−17.2	−13.0
% N in level	Far Below Basic	No schools	1.0 %	0.2 %	3.9 %	3.1 %	3.7 %	3.9 %	2.6 %	2.5 %
	Below Basic		6.5 %	3.3 %	22.2 %	20.4 %	18.9 %	22.2 %	17.2 %	13.6 %
	Basic		20.8 %	14.9 %	32.1 %	31.4 %	34.3 %	32.1 %	29.0 %	27.8 %
	Proficient		29.7 %	28.5 %	26.5 %	28.0 %	27.9 %	26.5 %	28.4 %	28.1 %
	Advanced		41.9 %	53.2 %	15.3 %	17.0 %	15.2 %	15.3 %	22.8 %	28.0 %

Levels and growth based on average CST scores (Math, ELA). Results similar based on Math or ELA individually. Results available upon request. High-SES schools<50 % and low-SES students, low-SES schools≥50 % low-SES students

Table 7 Correlations between incoming CST levels and school value-added scores

Previous year CST level	School value-added
% far below basic	−0.180
% below basic	−0.336*
% basic	−0.135
% proficient	−0.038
% advanced	0.256*

Four-year average correlations for *N*=29 schools

that are outside a school's control in order accurately predict a school's effectiveness (Creemers and Scheerens 1994; Creemers 2002; Sammons 2007). Additionally, data from several districts are needed in order to increase meaningful interpretation and allow generalizations to a wider population.

This study is a case study in California and the context in other districts or states can be quite different. However, California is an important case due to its racial and economic diversity. The changing demographics over the last two decades points to a trend of increased diversity in most states (Hobbs and Stoops 2002) and California is therefore an important case which could offer relevant insights for other states.

6 Discussion and conclusions

School accountability is a hotly debated topic in education. Several states and school districts in the USA have implemented or are planning implementation of value-added measures for schools or teachers. However, there is still considerable debate over the reliability and validity of VAMs as well as which model gives the best estimates (Braun 2005; Harris 2011; McCaffrey et al. 2004; Sanders et al. 2009). These questions are vital, as the choice of model affects the estimate and essentially determines if a teacher or school is labeled effective or not. Nevertheless, proponents of VAMs stress that the current model is inadequate as well (Harris 2011). Schools are penalized for the composition of their students and high-poverty schools face tremendous challenges in meeting the current accountability measures (Kim and Sunderman 2005). This study addressed one part of this debate comparing the currently used AYP and API metrics of school accountability in California with their respective value-added scores. The results support the critique voiced against current school accountability. API and AYP are very highly correlated with student background. Schools with a majority low-SES students lag behind schools from more affluent backgrounds and seem to struggle to reach the mandated goal of an API of 800. However, school value-added scores seem to measure something different. The relationship to background variables is far less compared with current metrics as the methodology controls for exactly such external factors of achievement. In other words, VAMs account for student background and therefore reduce the relationship between a school's value-added score and student background indicators. With a value-added approach, several low-SES schools showed remarkable performance, raising student's scores more than 200 points above expected levels. High-SES schools still performed well with very few schools performing below average.

In addition, this paper did not find any evidence that a school's value-added score is negatively impacted by high proportions of advanced students. This is important and could alleviate fears of schools that have shown high performance under API but are concerned with a ceiling effect of high-achieving students who cannot achieve large growth. This paper found little evidence of such a ceiling effect negatively influencing school value-added scores. However, future research should investigate this relationship in more detail.

This study holds clear implications for policy. Several schools in the district observed in this study did not reach the mandated levels of API. However, this study indicated that not all schools below an API of 800 are necessarily 'true' underperformers. In fact, some schools ranking at the bottom of the API scale showed remarkably high scores and raising student achievement more than expected. This study clearly exemplified the advantages of VAMs compared with the currently used metrics. With the high stakes associated with API and AYP and more importantly their high correlation with SES and race, these instruments could be equaled to what Peter Sacks called the "Volvo Effect" (Sacks 1999): One can guess a school's performance by looking at the type of cars in the school's parking lot. The value-added scores for schools however do not seem to commit this fallacy of strong dependency on student background. These results thus lend support to a school accountability model that incorporates value-added measures. A hybrid accountability system seems most beneficial which combines both API and value-added scores. Schools with high-achieving students and high API (e.g., above 800) are not in need of change. Schools with low API and high value-added should be rewarded for their above average performance given their context. Low performing schools on both scales would require support and additional evaluation efforts to determine necessary changes in order for students to receive adequate opportunity to learn and achieve at high levels. More research is certainly needed to investigate the questions addressed in this paper in more detail and with data from more than one district. However, based on the results of this study, a discussion on incorporating a value-added element in current school accountability measures seems relevant and necessary.

Acknowledgments The data for this study were collected under a grant to Claremont Graduate University (CGU), number ITQ-08-0502, NCLB Act of 2001, Title II-A, Improving Teacher Quality State Grants Program, administered by the California Postsecondary Education Commission. Key contributors to the study were Stacy Brown, Project Director at CGU and Pitzer College, and Wayne Snyder, Kristen Baldridge, Lorelei Coddington, Loris Fagioli, and Becky Orona at CGU.

Appendix

Table 8 Fixed effect model coefficients for 4 years in math and ELA

	Math				ELA			
	Year 2	Year 3	Year 4	Year 5	Year 2	Year 3	Year 4	Year 5
Previous achievement	0.67***	0.69***	0.67***	0.68***	0.69***	0.72***	0.70***	0.69***
Low SES	−0.78	0.57	−4.08***	−2.23	−2.00**	0.71	−2.29**	−0.57
SWD	2.33	−1.53	−2.05	5.39*	−8.49**	−4.85***	−2.96	1.49
Hispanic	8.50***	8.31***	12.65***	11.24***	3.29**	2.90*	6.65***	6.21***
Other	−7.12***	−7.92***	−8.35***	−5.30**	−4.31***	−3.92**	−5.38***	−4.32**
EL	−8.88***	−10.15***	−10.67***	−9.45***	−6.39***	−10.38***	−9.89***	−10.29***
Multigrade	−0.05**	0.09***	−0.05*	−0.04	−0.56***	−0.39***	−0.54***	−0.47***
Previous class achievement	−7.74***	6.38***	0.17	6.17***	0.29	3.67**	6.28***	2.52*
Class size	−1.31***	0.01	−0.80***	0.77**	−0.23	−0.54***	−0.10	−0.28
% low-SES	10.60	13.16	16.98**	9.61	−8.90	11.99*	16.30***	−38.86***
% SWD	20.05	−13.00	−7.76	−54.65***	−10.77	15.13*	−5.86	−16.72**
% Hispanic	8.49	10.79	−10.00	15.12	11.06	42.01***	38.61***	−21.86**
% other	−17.17	6.18	−95.04***	15.56	−11.19	−22.47**	−28.64***	4.84
% EL	−21.51***	34.26***	25.26***	40.58***	−66.86***	−49.79***	−65.33***	−39.49***
School size	3.06***	2.03***	3.61***	3.62***	0.15	0.15	0.42	0.71
Constant	−605.15**	−455.35***	−759.85***	−890.30***	313.64*	232.19***	208.53**	148.68
Observations	7,610	7,455	7,553	7,382	7,610	7,455	7,553	7,382
R-squared	0.61	0.61	0.58	0.58	0.64	0.60	0.60	0.60

***$p<0.01$; **$p<0.05$; *$p<0.1$

References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220.

Bill and Melinda Gates Foundation. (2010). Learning about teaching: Initial findings from the Measures of Effective Teaching project. Available from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

Blume, H. (2013). Deasy wants 30% of teacher evaluations based on test scores. Los Angeles Times. Available from http://articles.latimes.com/2013/feb/16/local/la-me-lausd-evals-20130216.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton: Educational Testing Service. Accessed 27 February 2008.

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. Los Angeles Times*. Boulder: National Education Policy Center.

Butrymowicz, S., & Garland, S. (2012). How New York City's value-added model compares to what other districts, states are doing. Hechinger Report. Available from http://hechingerreport.org/content/how-new-york-citys-value-added-model-compares-to-what-other-districts-states-are-doing_7757/.

California Department of Education (2013). Understanding the Academic Performance Index (API). Available from http://www.ed-data.k12.ca.us/Pages/UnderstandingTheAPI.aspx.

California Department of Education (2013). Academic Performance Index Reports: Information Guide. Available from http://www.cde.ca.gov/ta/ac/ay/documents/aypinfoguide11.pdf.

California Department of Education (2013). The public schools accountability act of 1999-CalEdFacts. Available from http://www.cde.ca.gov/ta/ac/pa/cefpsaa.asp.

California Department of Education (2013). Adequate Yearly Progress-CalEdFacts. Available from http://www.cde.ca.gov/ta/ac/ay/cefayp.asp.

California Department of Education (2013). Performance level tables for the California standards tests and the California alternate performance assessment. Available from http://www.ieminc.org/Assessment/starscaledscores.pdf.

Chudowsky, N., Braun, H. I., & Koenig, J. A. (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academy Press.

Cody, A. (2012). Educators issue VAM report for secretary Duncan. Education Week Teacher. Available from http://blogs.edweek.org/teachers/living-in-dialogue/2012/04/educators_issue_vam_report_for.html.

Creemers, B. P. (2002). The comprehensive model of educational effectiveness: Background, major assumptions and description. Accessed 22 January 2010.

Creemers, B. P., & Scheerens, J. (1994). Developments in the educational effectiveness research programme. *International Journal of Educational Research, 21*(2), 125–140.

Downes, D. M., & Vindurampulle, O. (2007). Value-added measures for school improvement. Education Policy and Research Division, Office for Education Policy and Innovation, Department of Education and Early Childhood Development.

Erpenbach, W. J., & Forte, E. (2008). *Statewide educational accountability systems under the NCLB Act: A report on 2008 amendments to state plans*. Washington, DC: Council of Chief State School Officers.

Ewing, J. (2011). Mathematical intimidation: Driven by the data. *Notices of the AMS, 58*(5), 667–673.

Goe, L. (2008). Key issue: Using value-added models to identify and support highly effective teachers. ETS. Available from http://hechingerreport.org/content/how-new-york-citys-value-added-model-compares-to-what-other-districts-states-are-doing_7757/.

Goldhaber, D. (2002). The mystery of good teaching: Surveying the evidence on student achievement and teacher's characteristics. *Education Next, 2*, 50–55.

Goldhaber, D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129–145.

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

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the Long-term stability of estimated teacher performance. *Economica, 80*(319), 589–612.

Goldhaber, D., Walch, J., & Gabele, B. (2012). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. CEDR Working Paper 2012-6, University of Washington, Seattle

Guarino, C., Reckase, M. D., & Wooldridge, J. (2012). Can value-added measures of teacher performance be trusted? IZA Discussion Paper No. 6602.

Hanushek, E. A. (2009). Teacher deselection. *Creating a New Teaching Profession, 168*, 172–173.

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

Hart, G. K., & Brownell, N. S. (2001). An Era of educational accountability in California. *The Clearing House, 74*(4), 183–186.

Hobbs, F., & Stoops, N. (2002). *US Census Bureau, Census 2000 special reports, Series CENSR-4, Demographic trends in the 20th century*. Washington, DC: US Government Printing Office.

Jakubowski, M. (2008). Implementing value-added models of school assessment. EUI Working Papers RSCAS 2008/06.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the "black box" of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*(3), 181–198.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3–13.

LA Times. (2011). Grading the teachers: Value-Added analysis. LA Times. Available from http://www.latimes.com/news/local/teachers-investigation/.

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*(4), 835–851.

Lee, J. (2003). Evaluating rural progress in mathematics achievement: Threats to the validity of" adequate yearly progress. *Journal of Research in Rural Education, 18*(2), 67–77.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education, 19*(1), 5–15.

Linn, R. L. (2008). *Validation of uses and interpretations of state assessments*. Washington, DC: Council of Chief State School Officers.

Manwaring, R. (2010). *Restructuring "restructuring": Improving interventions for low-performing schools and districts*. Washington, DC: Education Sector. 20.

McCaffrey, D. F., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability (Vol. 158)*. Santa Monica: Rand Corporation.

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.

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

Mintrop, H., & Trujillo, T. (2007). The practical relevance of accountability systems for school improvement: A descriptive analysis of California schools. *Educational Evaluation and Policy Analysis, 29*(4), 319–352.

NCLB (2001). No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425.

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

OECD. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD Publishing.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton: Educational Testing Service.

Ray, A., McCormack, T., & Evans, H. (2009). Value added in English schools. *Education Finance and Policy, 4*(4), 415–438.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rothstein, J. (2007). *Do value-added models add value? Tracking, fixed effects, and causal inference. Center for economic policy studies*. Princeton: Princeton University.

Rothstein, J. (2008). *Teacher quality in educational production: Tracking, decay, and student achievement*. Cambridge: National Bureau of Economic Research.

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

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge: Perseus Books.

Sammons, P. (2007). *School effectiveness and equity: Making connections*. Reading: CfBT.

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

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

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. Cary: SAS Institute Inc.

Saunders, L. (1999). A brief history of educational "value added": How did we get to where we are? *School Effectiveness and School Improvement, 10*(2), 233–256.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. Washington, DC: National Center for Education Evaluation and Regional Assistance. 64.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research, 6*(2), 103–114.

Strathdee, R., & Boustead, T. (2005). Measuring "value added" in New Zealand schools. *New Zealand Annual Review of Education, 14*, 59–76.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement, 22*(4), 393–413.

Wilson, S. F., & Floden, R. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education, 53*(3), 190–204.