The Random Assignment of Students Into Elementary Classrooms: Implications for Value-Added Analyses and Interpretations

Noelle A. Paufler and Audrey Amrein-Beardsley

Am Educ Res J published online 22 October 2013
DOI: 10.3102/0002831213508299

American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–35
DOI: 10.3102/0002831213508299

The Random Assignment of Students Into Elementary Classrooms: Implications for Value-Added Analyses and Interpretations

Noelle A. Paufler
Audrey Amrein-Beardsley
Arizona State University

Value-added models (VAMs) are used to measure changes in student achievement on large-scaled standardized test scores from year to year. When aggregated, VAM estimates are used to measure teacher effectiveness and hold teachers accountable for the value they purportedly add to or detract from student learning and achievement. In this study, researchers examined the extent to which purposeful (nonrandom) and random assignment of students into classrooms occurs in Arizona elementary schools (Grades 3–6). Researchers found that overwhelmingly, students are not randomly assigned and administrators, teachers, and parents play a prodigious role in the process. Findings have current implications for value-added analyses and the extent to which nonrandom assignment practices might impact or bias teachers' value-added scores.

Keywords: educational policy, teacher and school evaluation, high-stakes testing, accountability, value-added models, statistical models and assumptions

---

Noelle A. Paufler is a doctoral candidate studying educational policy and evaluation at Arizona State University, PO Box 37100, Phoenix, AZ 85069-7100; e-mail: *noelle.paufler@asu.edu*. Her research interests include research methods and educational policy and more specifically, the impact of value-added measures and their related teacher evaluation systems on educators, their professional practices, and the educational system as a whole.

Audrey Amrein-Beardsley, PhD, is an associate professor in Mary Lou Fulton Teachers College at Arizona State University. Her research interests include educational policy, educational measurement, and research methods and more specifically, high-stakes tests and value-added methodologies and systems.

Introduction

Given the heightened policy and pragmatic interest in value-added models (VAMs), attention has been given to the degree to which the purposeful (nonrandom) sorting of students into classrooms and schools matters and what this means for making reliable, valid, and unbiased estimates of the value-added by teachers and schools to student achievement. Researchers continue to demonstrate possible bias in value-added estimates (Capitol Hill Briefing, 2011; Dills & Mulholland, 2010; Hermann, Walsh, Isenberg, & Resch, 2013; Hill, Kapitula, & Umlan, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rothstein, 2009, 2010; Stacy, Guarino, Recklase, & Wooldridge, 2012). Researchers also continue to demonstrate that bias might occur more often when homogeneous sets of students (e.g., English language learners [ELLs], gifted and special education, racial minority, eligible for free or reduced lunches, retained in grade, in remedial programs) are purposefully placed in more concentrated numbers into some classrooms or schools than others (Baker et al., 2010; Capitol Hill Briefing, 2011; Goldhaber, Gabele, & Walch, 2012; McCaffrey, Lockwood, Koretz, Louis, & Hamilton 2004; Rothstein & Mathis, 2013). Accordingly, disputes about using VAM estimates for making high-stakes decisions about teacher and school quality have become increasingly relevant, with continuing concerns about whether biased estimates are being used to unfairly reward or penalize teachers and schools (Glazerman & Potamites, 2011; Hermann et al., 2013; Kersting, Chen, & Stigler, 2013; Raudenbush, 2004; Raudenbush & Jean, 2012).

The primary issue here is about whether even the most sophisticated VAMs can measure value-added in unbiased ways. Oftentimes, value-added statisticians assert that the achievement differences among students that occur due to nonrandom placement and sorting mechanisms can be controlled for, ultimately making nonrandomness a nonissue (Ballou, Sanders, & Wright, 2004; Goldhaber & Theobald, 2012; Meyer & Dokumaci, 2010; Sanders, 1998; Sanders & Horn, 1998; Wright, Horn, & Sanders, 1997). However, along with such an assertion come "heroic assumptions" (Rubin, Stuart, & Zanutto, 2004) that, although rarely discussed in the literature, researchers are gradually discrediting (Hermann et al., 2013; Koedel & Betts, 2007, 2010; Stacy et al., 2012). These assumptions include but are not limited to assumptions about linearity, manipulability, independence of errors, homoscedasticity, and most pertinent here, strongly ignorable student assignment (Braun, 2005; Reardon & Raudenbush, 2009; Scherrer, 2011).

Random Assignment

"Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions"

(Wilkinson & Task Force on Statistical Inference, 1999; see also Cook & Campbell, 1979). The purpose of random assignment is to make the probability of the occurrence of any observable differences among treatment groups (e.g., treatment or no treatment) equal at the outset of any experiment or study.

In this case, random assignment would involve using probabilistic methods to assign students to different treatment groups (e.g., classrooms or schools). This would help to ensure that the student characteristics that might bias treatment effects (e.g., different teacher- or school-level effects) are equally probable across comparison groups (e.g., students within classrooms with different teachers or students within different schools). This would help to make causal statements about treatment effects (e.g., teacher or school effects), using output indicators (e.g., growth in student achievement) more validly interpretable using standard statistical reasoning approaches. The *p*-value of any teacher or school effect would then accurately represent the probability of finding an effect at least as large as was found, simply due to chance differences in student characteristics. If the *p*-value is small, then there is evidence supporting the assertion that the teacher or school has a different impact on student scores than other teachers or schools.

As the ideal empirical approach for value-added analyses, students would be randomly assigned to classrooms such that any differences in value-added estimates would be attributable to the constructs being measured: teacher and school effects (Ballou, 2012; Ehlert, Koedel, Parsons, & Podgursky, 2012; Glazerman & Potamites, 2011). If all students could be randomly assigned to classrooms (and teachers to classrooms), the confidence with which we could make valid inferences using value-added scores would substantially increase (Corcoran, 2010; Guarino, Reckase, & Wooldridge, 2012; Newton et al., 2010; Rothstein, 2009). Random assignment would presumably mitigate the bias present without random assignment (Harris, 2009; Rothstein, 2010) and therefore help to control for the biasing effects of student background variables on value-added estimates (Ballou et al., 2004; Kupermintz, 2003; McCaffrey et al., 2004). However, it should be noted that if random assignment procedures were used to assign students to classrooms for accountability purposes, large class sizes would be necessary to ensure adequate power (Dunn, Kadane, & Garrow, 2003; Raudenbush, 2004). Even with class sizes large enough for adequate statistical power, it would still be possible to make Type 1 (rejection of a null hypothesis when the results can be attributed to chance) and Type 2 (failure to reject a null hypothesis when the results are not likely due to chance) errors when making inferences about teachers and/or schools depending on the magnitude of teacher and school effects. Guaranteeing adequate sample sizes, though, would not necessarily be in the best interests of students (Glass & Smith, 1979; Rivkin, Hanushek, & Kain, 2005; Word et al., 1990).

Notwithstanding, given the fact that value-added estimates are most often calculated when random assignment is not possible, under quasi-experimental conditions (Cook & Campbell, 1979), it is often necessary to assume that even if random assignment practices are not used, assignment practices are "effectively, if not formally, random" (Reardon & Raudenbush, 2009, p. 497). It must be assumed that any school is as likely as any other school, and any teacher is as likely as any other teacher, to be assigned any student who is as likely as any other student to have similar backgrounds, abilities, aptitudes, dispositions, motivations, and the like. This makes it more defensible to make more conservative statements about attribution, versus statements about direct causation (Ballou, 2012; Kersting et al., 2013; Raudenbush, 2004).

Assumptions About Random Assignment

A director of research and evaluation at the Arizona Department of Education (ADE) recently made public his or her assumption about random assignment—that, in general, students across the state were assigned randomly to classrooms. The director expressed this during a committee meeting regarding the VAM that state legislators were to adopt and implement, at the director's recommendation, as part of the state's new teacher accountability system. After sharing these beliefs, the director was criticized for making what was considered a highly false assumption. Attendees who opposed the director's comments argued instead that student assignments across the state were made in highly nonrandom ways. The director then noted the state needed more information about whether students were being purposefully (nonrandomly) assigned to classrooms. These statements inspired this study.

However, while the director's comments were made clear, and scrutinized as such, others have supported similar versions of this assumption. Harris (2011), for example, wrote (without supportive evidence) "in elementary schools, there is typically little tracking across classes (though teachers do track within classes)" (p. 114). He added that "in middle school, tracking is more common; and in high school, it is almost universal" (p. 114; see similar statements in Harris, 2009; Harris & Anderson, 2013). Rivkin and Ishii (2008) made an analogous assertion (without supporting evidence) that the systematic sorting of students "is much more prevalent in middle school than in the early grades," ultimately making attempts to produce unbiased value-added estimates "more difficult in middle school than in elementary school" (p. 17).

Guarino, Reckase, et al. (2012) claimed (without supportive evidence) that students are typically *not* randomly sorted into classrooms, but rather students are sorted using only students' perceived and actual academic abilities. They made explicit their assumptions about how tracking by academic

ability is the only existing student sorting mechanism and the one with which value-added statisticians should be primarily concerned (pp. 15–16). Hence, controlling for academically based student placement practices "explicitly controls for *the* [emphasis added] potential source of bias" (p. 30; see also Harris, 2009).

Elsewhere, many statisticians use VAMs that assume randomness, even though they do not necessarily believe that in actuality students are randomly assigned into classrooms. Instead, they claim that student assignments need not be made randomly if the most sophisticated models are used to estimate value-added, with controls that account for students' prior achievement(s) and sometimes other variables, as needed and when available. Put differently, this set of statisticians suggests that if student placements are nonrandom, complex statistics (e.g., student, classroom, and school fixed effects; nesting strategies that account for the nested structures in which students coexist; student-level covariates and other sophisticated controls; blocking, shrinkage, and bias compression strategies; ordinary least squares [OLS] estimators; etc.) can tolerably counter for the nonrandom effects that occur outside of experimental conditions.

Value-Added Research Center (VARC) model developers, for example, explicitly state that their advanced value-added model "produces estimates of school productivity—value-added indicators—under the counterfactual assumption that all schools serve the same group of students. This facilitates apples-and-apples school comparisons rather than apples-and-oranges comparisons" (Meyer & Dokumaci, 2010, p. 3). Developers of the SAS Education Value-Added Assessment System (EVAAS) note as well that their model is not biased by nonrandom student placement practices given the complex systems and controls they too have put into place (Sanders & Horn, 1998; Sanders, Wright, Rivers, & Leandro, 2009). They, and others, argue that the random assignment of students to teachers' classrooms, while ideal, is for the most part unnecessary (see also Bill & Melinda Gates Foundation, 2013; Glazerman & Potamites, 2011). The complex strategies and controls they use make the biasing effects of nonrandom placement practices effectively "ignorable" (Reardon & Raudenbush, 2009, p. 504). As for the errors that cannot be controlled, statisticians can use confidence intervals to better situate value-added estimates and their (often sizeable) random errors, so the errors can be better understood.

In sum, some assume that random assignment is not an issue in certain grade levels versus others, some assume that student sorting occurs using only students' prior academic achievements so it can be easily controlled, and some assume that, regardless, the statistical strategies and controls used are advanced enough to control for most if not all of the bias that might occur. These strategies and controls are discussed next.

Advanced Statistical Strategies and Controls

To reduce the errors often caused by bias, statisticians always control for at least one and preferably more years of students' previous test scores (e.g., using or exploiting covariates in lieu of randomized experimental conditions) to help adjust for the starting abilities of the students nonrandomly sorted into classrooms and schools. Most, if not all, agree that the most critical and most important VAM-based adjustment is students' prior achievement (Glazerman et al., 2011; Harris, 2009). As such, controlling for prior achievement helps to "level the playing field," so to speak, as this helps to ameliorate the biasing impact that extraneous variables have on achievement over time; although, whether this works is another source of contention (Ballou, 2012; Rothstein, 2009; Sanders et al., 2009).

Sanders et al. (2009), for example, argue that controlling for such extraneous variables is unnecessary because including students' prior test scores effectively controls for the extraneous variables intentionally excluded. This allows students to serve as their own controls (Ballou et al., 2004; Cody, McFarland, Moore, & Preston, 2010; Goldhaber & Theobald, 2012; Sanders, 1998). Additionally, Sanders et al. (2009) purport they have evidence that this works in that growth, when properly assessed using students' test prior scores, is not highly correlated with students' background variables (mainly race and poverty) using the EVAAS model. However, they do not provide statistical evidence of this assertion (e.g., correlations among levels of growth and race/poverty). Instead, they write, "correlations are modest at worst and essentially zero at best" (p. 6; see also Sanders, 1998), leaving "modest correlations" open to interpretation. Consequently, and regardless, they recommend adjustments for variables other than students' prior test scores not be made.

Researchers conducting secondary analyses of EVAAS data, however, have noted that bias still exists within their model, especially when highly homogenous sets of students (including large proportions of racial minority students) are not randomly assigned into classrooms (Kupermintz, 2003; see also Goldhaber et al., 2012; Guarino, Maxfield, Reckase, Thompson, & Wooldridge, 2012; Newton et al., 2010). Allowing students to serve as their own controls, in other words, only controls for bias in starting ability due to student traits. It does not, however, address the differential probabilities that unique students with diverse background characteristics (e.g., language proficiency, familial support, dissimilar motivations, access to technologies and resources outside of school, etc.) might otherwise have for making discrepant gains from year to year. This ultimately causes disparities both during the school year and over the summer months, especially considering that the pretests and posttests used to measure value-added encapsulate the summers. There still exists a lack of controls to directly counter for these effects (Baker et al., 2010; Corcoran, 2010; Gabriel & Allington, 2011; Harris, 2011).

That said, some statisticians integrate additional controls to account for some of these other, uncontrollable influences. They do this under the assumption that by controlling for some additional observable factors they might also control for other nonobservable factors, particularly those that might be more difficult or impossible to capture. Control variables most often incorporated include but are not limited to student-level variables (e.g., race, ethnicity, eligibility for free or reduced lunch prices as a crude proxy for socioeconomic backgrounds, ELL status, involvement in special education and gifted programs) as well as, sometimes, classroom- and school-level variables (e.g., daily attendance, prior teachers' residual effects, multiple teachers' fractional effects). The inclusion of combinations of additional controls facilitates better model adjustments, again to make such analyses less subject to bias (Braun, 2005; Goldhaber & Theobald, 2012; Hill et al., 2011; McCaffrey et al., 2004; Meyer & Dokumaci, 2010; Newton et al., 2010). What is still widely contested, however, is whether including these variables works to control for the other, nonobservable, yet potentially biasing variables as well.

Some researchers claim that the observable variables typically included within VAMs tend to be imperfect because they are very rudimentary proxies of the wider group of variables causing bias. So while including additional variables helps strengthen the models, this does not and might not ever control for or limit bias down to acceptable levels (Ehlert et al., 2012; Glazerman & Potamites, 2011; Kersting et al., 2013). They also question, for example, whether using binary (e.g., using ones and zeros) and categorical variables (e.g., using a range of numbers to represent categories without inherent values) can correctly capture students' ELL status, poverty-based realities, ranges of disabilities, and the like. Reducing highly complex phenomenon into mathematical representations and codes is much less exacting than it seems.

For example, while gender might be the most reasonable use of a dichotomous variable (i.e., male and female), of interest here is whether numeric categorizations of things like students' disabilities can effectively differentiate among and capture the learning trajectories of students who might be intellectually challenged, emotionally disabled, autistic, or a combination. Similarly, whether race can be isolated and effectively captured using categorical variables (e.g., 1 = American Indian, 2 = Asian American, 3 = African American, etc.) and whether these categories can be used to account for students' learning trajectories by race causes trepidations as well (see e.g., Briggs & Domingue, 2011). Issues with heteroscedasticity, or variation in the accuracy of predictions for students with lower prior achievement (e.g., students who also come from racial minority backgrounds, receive free or reduced lunches, are ELLs), also cause problems when producing unbiased estimates (Hermann et al., 2013; Stacy et al., 2012).

Nevertheless, value-added statisticians often employ the aforementioned variables, or a combination of those to which they have access, to

also implicitly control for the other nonobservable variables at play. The nonobservable variables not typically available in the large-scale data sets, and therefore not included when conducting value-added analyses, include but are not limited to students' behavior, discipline, or suspension records; students' self-handicapping and other dispositional, personality, motivational, attitudinal, or behavioral measures; students' family support systems including access to resources, books, and technologies within the home; students' parental supports, parents' levels of education, and parents' direct and indirect involvement in their children's learning; whether students attend summer school, access libraries or other public resources outside of school, or have access to tutors; and the like (Briggs & Domingue, 2011; Harris & Anderson, 2013; Rivkin & Ishii, 2008; Rubin et al., 2004).

Because all of these variables impact student learning over time, the question then becomes whether what is available and typically included can realistically account for all that is not. While the general rule here is to account for as many variables as possible (Harris, 2011), the factors that impact student achievement and growth over time but that are not included when estimating value-added still seem to be causing bias (Glazerman & Potamites, 2011; Newton et al., 2010; Rothstein, 2009, 2010). This is true even when the most sophisticated controls have been deployed (Capitol Hill Briefing, 2011; Guarino, Maxfield, et al., 2012; Guarino, Reckase, et al., 2012; Hermann et al., 2013; Koedel & Betts, 2007, 2010; Stacy et al., 2012).

What must be understood better is the extent to which student placement practices are effectively, if not formally, random; whether student achievement matters as much as assumed when students are assigned to classrooms; in what ways students are otherwise assigned to classrooms; and what all of this might mean for value-added. This is what researchers sought to do in this study—to understand how student assignment occurs in practice to help determine whether what researchers are typically accounting for *might* control for selection bias or make nonrandom assignment "ignorable" as often assumed (Reardon & Raudenbush, 2009, p. 504). Researchers did not set out to prove that VAM models are biased however.

Student Placement Practices

Whether students have been randomly assigned to schools and classrooms has not mattered much in the past because until recently, teachers were not typically held accountable for the test scores their students attained (i.e., once per year on traditional "snapshot" standardized tests). Most administrators who evaluated teachers, as well as the teachers themselves, realized that the nonrandom placement of students into classrooms could not be accounted for and thus should neither be used to gauge the teachers' effectiveness nor influence ratings of their performance. It did not make sense, for example, to penalize teachers whose classes were populated

with more "difficult to teach" students or to reward teachers of classes disproportionately filled with higher achievers.

But relatively limited research has been conducted to explore how students are assigned to classrooms in schools (Burns & Mason, 1995; Dills & Mulholland, 2010; Monk, 1987; Player, 2010; Praisner, 2003), and none of these studies have been conducted in the context of value-added. In chronological order, Monk (1987) found that the use of categories, often based on student demographic variables and previous academic performance, was the most common method used to assign students to classrooms. Burns and Mason (1995) concluded that principals of traditional or single-track schools had greater flexibility in creating heterogeneous classrooms based on students' ethnicity, gender, behaviors, language proficiency, parental requests, and previous interactions with teachers or other students. In multitrack schools, however, principals attempted to cluster students homogeneously. Praisner (2003) found that placement decisions, especially for students with disabilities, were largely affected by principals' attitudes, values, and professional coursework and training. Dills and Mulholland (2010) demonstrated issues with the ways students are placed into certain classes with certain class sizes using students' demographic variables (e.g., prior student behaviors). Player (2010) found that, especially as a form of nonmonetary compensation or benefit, principals had an incentive to assign higher achieving students with teachers they favored most and placed lower achieving students (e.g., males, students eligible for free or reduced lunches, students with disabilities) with teachers they favored less. This certainly has implications for value-added, but again in this context, while researchers are increasingly demonstrating that such nonrandom assignment practices continue to bias estimates, they are not necessarily evidencing how or why.

Most recently, and perhaps most notably, Mathematica Policy Research statisticians demonstrated that the VAM-based estimates for teachers who teach inordinate numbers of students with "harder-to-predict" achievement (i.e., students with relatively lower prior levels of achievement and who receive free or reduced lunch prices) are less precise, despite the sophisticated controls used (Hermann et al., 2013; McCaffrey, 2012). They also evidenced that the methods typically used to control for the nonrandom placement of students across most VAMs (e.g., shrinkage estimation methods like the Empirical Bayes approach) do not effectively work (Hermann et al., 2013; see also Guarino, Reckase, et al., 2012). This was also supported by Guarino, Maxfield, et al. (2012) who wrote that "although these estimators generally perform well under random assignment of teachers to classrooms, their performance generally suffers under non-random assignment when students are grouped based on prior achievement" (p. 1).

Ballou (2002) and Kupermintz (2003) demonstrated with the EVAAS that the blocking strategies statisticians use to "level the playing field" do not effectively block out bias either, whereas teachers in classrooms and schools

with highly homogenous and relatively higher racial minority populations still tended to exhibit lower value-added (see also Goldhaber et al., 2012; McCaffrey, 2012; McCaffrey et al., 2004; Stacy et al., 2012). Guarino, Reckase, et al. (2012) supported this as well, writing, "it is clear that every estimator has an Achilles heel (or more than one area of potential weakness)" (p. 15; see also Guarino, Maxfield, et al., 2012).

Hill et al. (2011) demonstrated that within-school sorting of higher achieving students into the classes of more effective teachers biased estimates even when the biasing variables were included in the models. While this was more evident among the more simplistic VAMs used (see also Newton et al., 2010), this also occurred when more sophisticated controls were employed. McCaffrey et al. (2004) demonstrated that the same teachers consistently demonstrated more effectiveness when they taught higher achieving students, fewer ELLs, and fewer students from low-income backgrounds. They concluded that "student characteristics are likely to confound estimated teacher effects when schools serve distinctly different populations" (p. 67; see also Baker et al., 2010; McCaffrey et al., 2004). Newton et al. (2010) found too that estimates were significantly and negatively correlated with whether teachers taught inordinate proportions of ELLs, racial minority students, and students from low-income backgrounds and inversely whether teachers taught inordinate proportions of students who were female in reading/language arts, tracked in mathematics, Asian American, and living with more educated parents.

Also notably, Rothstein (2009, 2010) demonstrated that given nonrandom student placement (and tracking) practices, value-added estimates of future teachers could be used to predict students' past levels of achievement (counterintuitively). This suggests that students are systematically grouped in ways that bias value-added estimates and that systematic student assignment and sorting practices are far from random. Otherwise, such backwards predictions could not have been made (see also Briggs & Domingue, 2011; Koedel & Betts, 2010).

Stacey et al. (2012) also evidenced that students from low socioeconomic backgrounds and students with relatively lower levels of past achievement yielded less accurate teacher-level value-added estimates than their more advantaged peers. Teachers with students nonrandomly assigned to their classes, therefore, "might be differentially likely to be the recipient of negative or positive sanctions . . . and more likely to see their estimates vary from year to year due to low stability" (p. 1; see also Hermann et al., 2013).

Otherwise, only two studies thus far have actually used randomized experimental methods to test for selection bias (Bill & Melinda Gates Foundation, 2013; Kane & Staiger, 2008). Kane and Staiger (2008) asserted that accounting for students' achievement histories was sufficient to control for selection bias or selection on nonobservable variables. However, the

definition of sufficient was left open to interpretation; the value-added estimates produced also used the aforementioned Empirical Bayes methods (Hermann et al., 2013; Stacy et al., 2012; see also Chetty, Friedman, & Rockoff, 2011), and study results applied to a fairly narrow sample of participating and compliant schools (Guarino, Reckase, et al., 2012; Harris & Anderson, 2013). More importantly, results were not based on a true randomized experiment whereby students were not randomly assigned to classrooms. Instead, "principals in each of the schools were asked to draw up two classrooms they would be equally happy to have assigned to each of the teachers in the pair [or classroom dyads]. The school district office then randomly assigned the [dyad] classrooms to the [two] teachers" (Kane & Staiger, 2008, p. 2).

Related, in the final of $45 million worth of Bill & Melinda Gates Foundation Measures of Effective Teaching (MET) studies (2013), a majority of schools and teachers reneged on their agreements to safeguard and follow through with the randomized design. This too impeded on the validity of findings (Rothstein & Mathis, 2013), although statisticians noted that they could control for this attrition as well (Bill & Melinda Gates Foundation, 2013).

Nonetheless, the majority of the aforementioned researchers have evidenced that the value-added estimates of teachers who teach largely homogenous groups of students, students who are often nonrandomly sorted into classrooms, and despite the sophistication of the statistical controls used to eliminate bias, are still biased. As major studies continue to evidence that nonrandom sorting practices complicate estimates, value-added researchers must continue to acknowledge that this is an issue and that this issue deserves even more serious attention.

Purpose of the Study

In this study, researchers investigated the methods that elementary school principals in Arizona typically use to assign students to teachers' classrooms. Again, some things are known regarding how students are placed in general, but this is now much more important to explicate and understand because of the new accountability initiatives and value-added systems being adopted across the nation. This is the first study to provide evidence about the extent to which the purposeful (nonrandom) and random assignment of students into classrooms occurs in the context of the value-added.

The purpose of this study was to add to our collective thinking in this area, again given potential implications for making and better understanding value-added inferences and their evidence of validity. The three main research questions researchers addressed were the following:

Research Question 1: What are the methods elementary school principals typically use to assign students to teachers' classrooms?

Research Question 2: What are the key criteria elementary school principals typically use to place students if nonrandom practices are employed, and do these key criteria correlate with one another (i.e., are principals who use one academic indicator likely to also consider behavioral records to make placement decisions)?

Research Question 3: To what extent do students, teachers, and parents play a role in the student assignment process if nonrandom practices are employed?

Findings in this study were used to draw implications regarding how such practices *might* impact value-added inferences. While this is also the first study to provide concrete evidence that students are not randomly assigned to teachers, again, the question of whether certain placement practices really biased value-added estimates under varying conditions was not directly examined or explored. While it is certainly reasonable to ask to what extent bias occurs given varying student placement practices, in the state of Arizona, students' growth scores are still not linked to teachers' records to permit teacher-level value-added analyses for such a purpose. If this is done, it is only done at the district level. That said, it was impossible for researchers (barring collecting and combining the state's 227 districts', not including charter schools', files) to analyze, for example, whether the schools for which certain principals reported using certain student assignment techniques had more or less biased value-added scores. While this would have offered another major contribution to the VAM-based literature, such a study was well beyond the scope of this study, as well as beyond the data collection capacities of study researchers.

Instead, researchers addressed how the student assignment practices that were revealed might impact value-added estimates and inferences. This was done as situated in the previously stated assumptions value-added statisticians often make when controlling for observable variables, such as prior student achievement and student background variables, as well as those often used as proxies for nonobservables also at play. Such nonobservable variables may include: students' levels of motivations, aptitudes, dispositions; parental support systems; access to technologies and resources outside of school; and the like.

The fundamental question here is whether the nonrandom student assignment practices discovered in this study might logically lead to biased VAM estimates; that is, if the nonrandom student sorting practices go beyond that which is typically controlled for in most VAM models (e.g., academic achievement and prior demonstrated abilities, special education status, ELL status, gender, giftedness). For example, if behavior is used to sort students into classrooms but student behaviors are not typically accounted for in most VAMs, it is reasonable to assert that sorting students by their past and, related, potential behaviors could indeed cause bias.

This would make it reasonable to suspect that assignment using student attributes outside the scope of the data sets typically used to calculate value-added might bias estimates, even if other sophisticated controls are in place. That said, results from this, again the first large-scale study to survey elementary school principals regarding their student assignment practices in the context of value-added, should help us begin to better understand for what VAMs might control well and for what VAMs might not and might not ever control at all.

Research Methods

Survey-Research Design

In this survey-research study, researchers used a mixed-methods design to examine the use of purposeful (nonrandom) and random placement of students into classrooms in the public and charter elementary schools in the state of Arizona. Researchers designed a mixed-methods study in order to understand the different aspects of this complex social phenomenon better (Greene, 2007; Teddlie & Tashakkori, 2006), namely, the methods used for student assignments in elementary schools in Grades 3 through 6.

Specifically, researchers developed a survey (see the Appendix in the online journal) that contains concurrent quantitative and qualitative items; whereas quantitative responses could be numerically described, qualitative responses could help to inform and explain the quantitative data gathered, and qualitative responses could be converted to quantitative indicators to display and illustrate frequencies and trends (Greene, 2007; Teddlie & Tashakkori, 2006). This design allowed researchers to examine the data gathered for complementarity to strengthen assertions through the convergence of findings. This approach ultimately increased the strength of study results (Greene, 2007; Johnson & Onwuegbuzie, 2004; Teddlie & Tashakkori, 2006).

Participant Sample

Researchers invited elementary school principals from across Arizona to participate in an online survey. They obtained the contact information of public and charter elementary school principals (n = 2,447) from the Arizona Department of Education. They then removed the names of any principals for whom no email address was available, and they removed the names of all principals who oversaw nontraditional schools (e.g., alternative, special education, and vocational), primary schools (Grades Pre-Kindergarten–2), middle schools (Grades 7–8), and schools in which no students were currently enrolled. The final list included 1,265 principals (of the original 2,447 = 51.2%). All principals who remained in the final sample were emailed the survey instrument with an electronic invitation to respond via Survey Monkey.

Researchers kept the survey instrument open for just over 3 weeks in the spring of 2012. Using a confidence interval calculator with a 95% confidence level (±5% standard error), researchers determined that the sample of principals who responded during that time frame (n = 378/1,265, 30.0%) was large enough for sufficient power to be achieved and to draw conclusions using standard statistical approaches. While the sample of respondents was large enough to draw conclusions given statistical power and acceptable/low levels of standard errors, without representativeness ensured, researchers still cannot make a strong case that the results generalize regarding similar settings, people, or other like samples (Creswell, 2003; Ercikan & Wolff-Michael, 2009). There were also issues regarding potential response bias in that respondents may have differed from principals who were invited to participate but chose not to.

Based on recommendations made by Wilkinson and the Task Force on Statistical Inference (1999), researchers added evidence to support sample representativeness on a few other key (yet also imperfect) indicators that could be used to compare sample to population characteristics (see also Thompson, 2000). While the results of this study may not be generalizable to other similar samples, readers might make naturalistic generalizations from the findings within their own contexts and given their own experiences (Stake & Trumbull, 1982). Readers could gain general insights about student placement practices in elementary schools and what this might mean for value-added ratings, particularly given the experiences described by participants in this study.

Survey Instrument

Researchers dedicated nearly 6 months to the development of the survey. This included two pilot phases, during which researchers asked eight current or former principals in the state, who were not included in the final list of participants, to provide feedback regarding the survey format, length, and/or to provide comments or suggestions about specific sections, questions, or areas that researchers may not have included in the instrument. The final survey included 34 Likert-type and open-ended questions that researchers organized into six sections, each of which had a different focus: Section 1 included school demographic questions; Section 2 included questions about the responding principal and his or her pre-professional and professional training; Section 3 included questions about the methods the principal used to assign students to classrooms; Section 4 included questions about the role(s) teachers play, if any, in assignment decisions; Section 5 included questions about the role(s) parents play, if any, in their children's placements; and Section 6 included overarching questions (e.g., a question requesting principals' opinions about randomly assigning students to better measure value-added).

Data Analysis

For all participant responses (n = 378/1,265, 30.0%), researchers calculated descriptive statistics (means and standard deviations) using respondents' numerical responses to the sets and series of Likert-type items included in the survey instrument. Researchers then rank-ordered participant responses to demonstrate frequency and for descriptive purposes regarding school, student, and principal background variables. Researchers also calculated Pearson bivariate correlations among the key criteria that principals reported using to make placement decisions, noting statistically significant coefficients all the while ($p \leq .05$).

Researchers analyzed all qualitative data following the key methods and concepts of grounded theory (Glaser & Strauss, 1967; Strauss & Corbin, 1995), engaging in three rounds of "constant comparison" (Glaser & Strauss, 1967) and using a code-calculation spreadsheet to quantify the qualitative data for reduction, description, and conclusion-drawing purposes (Miles & Huberman, 1994). Specifically, researchers first analyzed the raw data for each open-ended question to identify instances or basic units of analysis. After inductively constructing working themes, they returned to the raw data to determine the frequency with which instances or units of analysis appeared in each set of responses (Erickson, 1986). They then collapsed the code clusters into a series of major and minor themes and quantified and ordered these from most to least often reported for comprehension and ease. To discover key linkages (Erickson, 1986) that existed between both the quantitative and qualitative data, researchers reviewed the entire corpus of data several times before drawing and substantiating final conclusions (Erickson, 1986).

Demographics and Sample Representativeness

As mentioned, while the response rate was not necessarily of concern, sample representativeness likely was. It was possible that principals who participated in this study were distinctly different from the principals who declined. Because this threatened both the validity and generalizations that could be made from study results, researchers further examined whether participants represented the characteristics of the general population of principals from which the sample came (Wilkinson & Task Force on Statistical Inference, 1999; see also Thompson, 2000).

Unfortunately, the data set to which researchers had access did not include any background or demographic variables about principals who were invited to participate. Only the email addresses of all elementary school principals in the state were included. Therefore, it was not possible to run statistical tests to examine the sample to general principal population characteristics, test for homogeneity, or test for significant likenesses or

differences among responding and nonresponding principals (e.g., using chi-square analyses).

Instead, researchers examined sample representativeness using logical and comparative yet nonstatistical approaches (Wilkinson & Task Force on Statistical Inference, 1999; see also Thompson, 2000). Researchers used the most current state-, county-, and school-level data available via the National Center for Education Statistics (NCES), the U.S. Census Bureau, and other local sources to examine sample-to-population characteristics, to help reduce or eliminate some of these potentially biasing elements. In one case, researchers were able to examine principals' years of experience as compared to the state population of principals, but otherwise, state-level information that matched the self-report data collected was not available for comparative purposes. These data are presented alongside sample demographics next.

School Size and Location. Principals who responded reported representing public and charter elementary schools of various sizes across Arizona that enrolled students in Grades 3 through 6. In terms of size, 78.9% (n = 291/369) enrolled more than 400 students. NCES data indicate that the average elementary school in Arizona enrolls 511 students (U.S. Department of Education, NCES, Common Core of Data [CCD], 2009–2010c), which indicates that enrollment in the schools of participating principals were of similar size to the average enrollment of elementary schools in the state.

In terms of location, respondents represented schools that were spread evenly across rural, urban, and suburban localities (rural n = 110/368, 29.9%; urban n = 120/368, 32.6%; and suburban n = 138/368, 37.5%). The vast majority of schools (n = 228/367, 62.1%) were located in Maricopa County. According to the U.S. Census Bureau (2011), this county is the largest in the state with 3.9 million people, out of the state's total population of 6.5 million (60.0%) currently residing in Maricopa County. Otherwise, principals from schools in 13 other counties also responded, most often representing Pima County (n = 34/367, 9.3%), Yuma County (n = 23/367, 6.3%), and Pinal County (n = 21/367, 5.7%). These three counties represent some of the highest populated counties next to Maricopa County according to the U.S. Census Bureau (2011). These data show that the participating principals represent the general population from all the major counties proportionately.

Student-Level Demographics. Survey results indicated that the schools from which principal respondents came had diverse student populations. More than 60.0% of students enrolled in almost half (n = 161/367, 43.9%) of the schools represented by the participating principals were from racial/ethnic minority backgrounds. This was also consistent with NCES data, which show that 42.3% of students in all Arizona public elementary schools are from racial minority backgrounds (U.S. Department of

Education, NCES, CCD, 1999–2000 and 2009–2010). While almost half of respondents served as principals in schools with more racially diverse student populations than the state average, the slight majority of other respondents did not.

In addition, findings indicated that 57.8% of the schools (n = 212/367) had an ELL population of 20.0% or less. This, too, was consistent with state-level data included as per the Office of the Auditor General (2007), although it is important to note that 10.9% of the schools represented in the study (n = 40/367) had a reported ELL student population of more than 60.0%.

Finally, NCES data regarding students eligible for free or reduced lunch prices also matched the demographics of the student populations in the responding principals' schools. Almost 60% of principals (n = 215/368, 58.4%) noted that more than 60% of students at their schools were eligible for this federal program, whereas 47.3% of all students in the state in 2009–2010 were eligible (U.S. Department of Education, NCES, CCD, 2009–2010a). While over half of respondents served as principals in schools with more eligible students than the state average, the remaining principals did not.

Researchers also verified the representativeness of the sample by the number of special needs students (U.S. Department of Education, NCES, CCD, 2009–2010b), gifted or talented students (U.S. Department of Education, Office for Civil Rights, 2004, 2006), and using other Title 1 data (U.S. Department of Education, NCES, CCD, 2010–2011). Gender demographics were not available.

Teacher and School Quality. In terms of the teaching staff at each of the represented schools, 61.1% of responding principals (n = 225/368) indicated that their school employed 26 to 50 teachers. At 86.7% of the schools (n = 313/361), 81.0% or more teachers were highly qualified in the subject area(s) that they taught (e.g., math, language arts, science, and social studies). Similar state reports show that only 1.7% of core academic courses statewide are not taught by highly qualified teachers (State of Arizona Department of Education, 2010–2011).

In terms of the ratings of the represented schools, 17.6% of the schools were rated as Excelling (n = 65/369), 16.5% were rated as Highly Performing (n = 61/369), and 48.2% were rated as Performing Plus (n = 178/369) in 2011 as per the state's No Child Left Behind (NCLB) Adequate Yearly Progress (AYP) requirements (State of Arizona Department of Education, 2010–2011). This correlated with state data, in that for the 2010–2011 school year 16% of all state schools were rated as Excelling, 14% were rated as Highly Performing, and 39% were rated as Performing Plus (State of Arizona Department of Education, 2010–2011).

Principal-Level Demographics. In terms of the responding principals' levels of advanced training, nearly all respondents had a considerable

amount of advanced training (e.g., certificates and/or master's or doctoral degrees) as well as extensive experience as administrators. When asked to describe their advanced training, almost all principals (n = 333/350, 95.1%) reported having earned a graduate degree. Of these, 58 respondents (n = 58/333, 17.4%) reported also having earned a doctoral degree.

In terms of principals' years of experience, survey results showed that most responding principals (n = 313/366, 85.5%) had more than 3 years of administrative experience. Notably, 26.0% (n = 95/366) reported having at least 13 years of administrative experience. NCES data supported this finding as well. Almost half (n = 168/367, 45.8%) of respondents reported holding their current position for 3 years or less, and as per the NCES data 57.8% of state principals reported the same. In addition, 13.1% (n = 48/367) of respondents reported serving at their school for at least 10 years, and as per the NCES data, 12.2% of principals statewide reported the same (U.S. Department of Education, NCES, Schools and Staffing Survey, 2007–2008).

These data should help to verify that respondents are representative of the statewide elementary school principal population. However, the sample is still limited; therefore, it is not possible to make other than naturalistic generalizations (Stake & Trumbull, 1982).

Results

Informed Placement Practices

Despite the aforementioned levels of advanced training and prior administrative experience, most principal respondents noted that the assignment of students was not discussed during their professional or administrative coursework (n = 284/363, 78.2%) or during any other professional development they had received since (n = 239/361, 66.2%). Those who recalled discussing the topic during coursework (n = 71/363, 19.6%) described the nature and extent to which the topic of the assignment of students was addressed. Most respondents noted that what was emphasized was the need to consider student background characteristics during the assignment process, most frequently citing the importance of making placement decisions using students' special education needs (n = 17/71, 23.9%), academic achievement or abilities (n = 15/71, 21.1%), gender (n = 9/71, 12.7%), and giftedness (n = 8/71, 11.3%), in that order.

Principal respondents also described discussing the importance of purposefully creating "balanced" and "heterogeneous" classrooms (n = 23/71, 32.4%). Those who discussed the assignment of students as part of other professional development activities (n = 105/361, 29.1%) mentioned similar topics, most often noting the importance of student background characteristics in the assignment process, namely, focusing again on language proficiency (n = 17/105, 16.2%), giftedness (n = 15/105, 14.3%), special education

needs (n = 14/105, 13.3%), and academic achievement (n = 14/105, 13.3%). Most principals (n = 308/353, 87.3%) noted that their district policy manual did not prescribe or mention a procedure for placing students into classrooms.

Methods of Assignment

When respondents described the various methods they used to assign students to classrooms in their schools, nearly all (n = 335/342, 98.0%) described procedures whereby administrators and teachers considered a variety of student background characteristics and student interactions to make placement decisions. In fact, in 98.0% (n = 335/342) of respondents' schools, random assignment to classrooms is not the general practice. Few (n = 25/342, 7.3%) principals mentioned the term *random* in their responses about their assignment practices at all.

In addition to students' academic achievement or ability (n = 188/342, 60.0%), behavior (n = 162/342, 47.4%), and special education needs (n = 147/342, 43.0%), principals frequently cited, as important considerations, the following in their open-ended responses: gender (n = 122/342, 35.6%), large-scaled standardized test scores (n = 98/342, 28.7%), and giftedness (n = 95/342, 28%). Very few principals (n = 34/342, 9.9%) identified students' racial or ethnic backgrounds as a factor in the placement process. Even fewer (n = 11/342, 3.2%) reported considering students' socioeconomic status when making placement decisions.

Given the set of Likert-type items used to identify the student characteristics considered when they are placed in classes, participants' open-ended responses matched or validated the quantitative findings. Principals reported the following in order of importance when making such placement decisions: students' prior academic achievement (M = 3.63, SD = 0.72), students' prior behavioral issues (M = 3.28, SD = 0.87), students' language status and/or levels of proficiency (M = 3.19, SD = 1.00), and students' perceived behavioral needs (M = 3.17, SD = 0.85). See Table 1 for all other student characteristics that principals reported, in order of most to least important with corresponding means and standard deviations.

For both exploratory and cross-validation purposes, researchers found statistically significant Pearson bivariate correlations between student background variables that principals reported using when placing students. Principals who considered students' prior behavioral issues were most likely to consider students' behavioral needs (r = .69, $p \leq .01$) to predict students' "best" placements. These principals were also most likely to examine students' discipline records (r = .68, $p \leq .01$) when making placement decisions. Principals who used large-scaled standardized test scores also considered district test scores when assigning students (r = .60, $p \leq .01$). These principals considered students' prior academic achievement in conjunction

Table 1

Student Background Characteristics Reportedly Considered by Principals when Assigning Students to Classrooms

	M	*SD*
1. Prior academic achievement	3.63	0.72
2. Prior behavioral issues	3.28	0.87
3. Language status and/or proficiency	3.19	1.00
4. Perceived behavioral needs	3.17	0.85
5. Prior grades	3.02	0.97
6. Prior large-scale standardized scores	3.02	1.01
7. Prior district test scores	2.94	1.03
8. Disciplinary records	2.87	1.02
9. Racial or ethnic backgrounds	1.76	1.08
10. Rates of absenteeism	1.62	0.87
11. Rates of attrition/transience	1.52	0.83
12. Socioeconomic backgrounds	1.32	0.71

Note. Likert items were scaled as follows: *strongly considered* = 4, *somewhat considered* = 3, *minimally considered* = 2, *not at all considered* = 1.

with preceding grades during the placement process ($r = .51$, $p \leq .01$) as well. Here, correlation coefficients helped to validate the responses from principals regarding the student characteristics that they considered during the placement process. See Table 2 for all correlation coefficients with levels of statistical significance.

Some principals ($n = 47/263$, 17.9%) also identified students' interactions with other students and teachers as critical factors that influence the placement process. These principals indicated that interactions among students, whether positive or negative, significantly impacted the learning environment in classrooms and also played a significant role in determining where students were placed for the following school year. For example, one principal described a common practice, explaining that teachers provide "information on behavior issues and/or students that should be placed in separate classes. Anything useful that can assist in the best placement for their students into the next grade" was reported as being important. Another principal added to this by expressing that a student's interactions with his or her peers can dramatically change the classroom dynamic, and "if students do have some behavior issues when they are with their peers, they may be assigned to separate classes the next year." Some principals ($n = 50/306$, 16.3%) mentioned negative interactions with other students as legitimate reasons to honor a parent's request for a specific placement for his or her child.

In addition, principals stated that placements often depended on students' learning styles and personality/compatibility characteristics ($n =$

Table 2

Pearson Correlation Coefficients Representing the Relationships Between Student Background Characteristics Reportedly Considered by Principals When Assigning Students to Classrooms

	1	2	3	4	5	6	7	8	9	10	11	12
1. Prior academic achievement	1	.33**	.17**	.37**	.51**	.40**	.43**	.28**	.08	.12*	.12*	.07
2. Prior behavioral issues		1	.27**	.69**	.23**	.07	.24**	.68**	.22**	.16**	.13*	.13*
3. Language status and/or proficiency			1	.31**	.07	.12*	.14*	.27**	.24**	.20**	.22**	.21**
4. Perceived behavioral needs				1	.24**	.15**	.23**	.59**	.26**	.22**	.20**	.19**
5. Prior grades					1	.35**	.37**	.25**	.16**	.12*	.12*	.16**
6. Prior large-scale standardized scores						1	.60**	.15**	.07	.34**	.30**	.19**
7. Prior district test scores							1	.25**	.11	.25**	.25**	.20**
8. Disciplinary records								1	.19**	.28**	.28**	.18**
9. Racial or ethnic backgrounds									1	.22**	.27**	.50**
10. Rates of absenteeism										1	.77**	.42**
11. Rates of attrition/transience											1	.49**
12. Socioeconomic backgrounds												1

*$p \leq .05$. **$p \leq .01$.

128/306, 41.8%). These decisions were most often informed by the comments and recommendations made by teachers, in addition to students' prior interactions with their teachers, their teachers' personalities, and their teachers' varying instructional and management styles. Thus, principals reported that they relied on teachers to make recommendations about student placements based on how students responded to them as instructors in the past. Some principals (n = 36/306, 11.8%) reportedly preferred placing a student with a "particular teacher that [a prior teacher felt] the child would be most successful with," and almost one-third of respondents (n = 92/306, 30.1%) reported placing students with teachers "who would best fit [the] learning needs [of individual students]." Principals also noted that teachers "assist [in] matching student personalities to teacher personalities" because "students may relate better to a specific teacher." One principal explained:

> Teachers are asked to build class loads. We do this to balance out the character of the teacher and students. We believe that this is helpful to have the best model to help kids do well within a class setting. Sometimes we get it wrong, and we change it as needed.

Finally, principals described other placement procedures, some of which created more heterogeneous classes and others creating homogeneity. In an effort to balance classes, principals reported using student information cards, for example, in efforts to create heterogeneous class lists. Other principals described the use of prescribed cluster grouping models when placing students, a practice that would yield more homogeneous groupings. Both of these methods often involved teachers and other school staff members.

The Roles of Teachers in the Placement Process

Related, results indicate that teachers are highly involved in the placement process in 88.7% of schools (n = 314/354). It is important to note that many principals specifically expressed their confidence in their teaching staff, describing their teachers as the best equipped to make such student placement decisions. A principal captured this by writing:

> They [the teachers] may know which teacher would be the best fit for their student moving to the next grade level. They also know which students may/may not work well together in the same class. They know the student best in the educational setting.

Respondents (n = 93/263, 35.4%) also described the role(s) played by teachers. Principals (n = 93/263, 35.4%) provided detailed descriptions of the role that teachers and other staff typically play in the placement process, where teachers, working individually or collaboratively within grade-level teams, were those charged with creating preliminary class lists based on

student background characteristics. According to respondents, teachers most often considered at least one of the following student characteristics: interactions with other students (n = 29/263, 11.0%), students' levels of academic achievement (n = 27/263, 10.3%), behaviors (n = 27/263, 10.3%), special needs (n = 27/263, 10.3%), and/or interactions with previous teachers (n = 18/263, 6.8%) when making placement recommendations. When placing students with the assistance of teachers, principals often reviewed these lists to make changes as needed. Principals never reported that it was solely the teachers' responsibility to make placement decisions.

Principals also frequently reported that they provided specific guidelines for teachers, directing them to create heterogeneous classrooms using most often (i.e., next to peer interactions) student characteristics that were academically related (n = 27/263, 10.3%). Some principals (n = 58/342, 17.0%), however, also required teachers to use cluster-grouping models to place gifted students, thus effectively creating more homogeneous classes. While principals often sought to "balance" classrooms as much as possible, in an effort to create more heterogeneous, harmonious learning environments, there were noteworthy exceptions.

Survey results also showed that teachers currently working in a school are asked to assist in matching the learning needs of students with future teachers' instructional styles, personalities, and other strengths. In describing this critical aspect of the placement process, one principal explained that teachers "complete a paper on each student . . . [and] list which, if any, particular teachers they feel the child would be most successful with and why." Another respondent noted the importance of teachers providing "learning modality information about students that helps in assigning students to match teaching strengths of teachers."

Some principals (n = 21/263, 8.0%) also described procedures where current teachers purposefully assign students based on their learning styles, namely, those who will likely benefit from a particular instructional or classroom management style. For example, one principal wrote that he or she has "some teachers who have strengths regarding language who can more effectively work with some demographics or even request [to] work with some of the most in need." Another responded, "The previous year's teachers place the students according to special needs, ELL, behavior, and levels of academics into the next [classroom with] A Teacher, B Teacher, C Teacher, D Teacher, or E Teacher." When reviewing these preliminary assignments again, however, several principals again noted the need for "balanced" (n = 95/342, 27.8%), "equal" (n = 29/342, 8.5%), "heterogeneous" (n = 25/342, 7.3%), and "fair" (n = 19/342, 5.6%) class lists.

One noted, for example, that he or she worked to ensure that "no classes are 'stacked' for a particular teacher . . . [and that] anyone could be assigned to any group." By making changes to class lists as needed, respondents frequently suggested that mismatched placements could be remedied

before the school year began. One principal explained that "lists are then given to the principal for final review with [the] school effectiveness mentor, counselor, [and] special needs representative." Again, while teachers may play a large role in the placement process, it was never reported that it was the sole responsibility of teachers to make placement decisions.

Citing the important role of other school staff members, a respondent wrote that "the school special education team (psychologist, social worker, speech pathologist, reading specialist, resource teachers, gifted teacher, and the principal) meet[s] and place[s] students who are on IEP's, 504 plans, behavior plans, in the gifted program, and those receiving special reading services." Notably, 40 respondents (n = 40/342, 11.7%) emphasized the need for input from others, including special education and special area teachers as well as support staff.

Lastly, and perhaps most importantly, by matching students with teachers in ways that principals hoped would create the best outcomes, again in terms of teachers' instructional styles, personalities, and/or other pedagogical strengths, principals rejected the idea of random assignment. The majority of respondents (n = 218/321, 67.9%) suggested that the use of random methods of assignment would be nonsensical, even if random assignment produced more valid value-added estimates. Principals thought that random assignment would be inappropriate and even harmful to students because this would not be in the children's best educational and developmental interests. Principals' actions during the placement process supported their collective argument as well. This will be discussed in more detail in an upcoming section.

The Role of Parents in the Placement Process

Sixty percent (n = 212/354, 59.9%) of responding principals noted that parents request specific placements for their children, and more than one-third of principals (n = 102/299, 34.1%) noted that they honored more than 80% of such requests. When asked to describe the circumstances under which the principal would consider the parents' requests legitimate, principals frequently cited requests based on students' learning styles, interactions with peers, or prior negative experiences between parents and teachers.

In terms of students' learning styles, some principals (n = 66/306, 21.6%) noted that they would attempt to honor a parent request regarding the best learning environment for his or her child. One principal described such a situation as follows:

> I will always meet and discuss [placements] with parents at their request. Occasionally the request is driven by a medical need or an IEP need. We attempt to provide input to the process for parents via parent input forms though there is a low rate of completion . . . as [the form] does not specify [a] teacher but rather [the] learning

> needs of the child. I'd always consider a request for a type of assignment . . . though we do not entertain requests for particular teachers.

Here, principals (n = 58/306, 19.0%) also cited prior negative interactions as a result of placements of siblings or relatives as legitimate reasons to honor specific requests. One principal explained that he or she would move a student to another class if unable to "remediate [the] problem between [the] parent and teacher even after [a] discussion [as a result of] a previous problem with the teacher with an older sibling." While a few principals (n = 13/306, 4.2%) expressed a willingness to make a placement change under such circumstances, they also expressed their desire to attempt to resolve any issues prior to moving the student. For example, a principal explained his or her response to such requests:

> Once teacher assignments are made, I typically have 8-10 change requests from parents. I meet with the parent and listen to their concern. Typically, I require the parent to try the assigned teacher. If after a two-week trial period, the concern remains, we meet with the teacher and try to resolve the issue within the classroom. If the issue then remains unresolved, I make a classroom change.

Another principal explained that,

> Current-year teachers supply the information used to balance out the classes. Teachers of the incoming classes only have input regarding students of families with whom they have had prior negative experiences. Avoiding situations that are predestined for problems is much easier before the classroom assignment has been made.

Another principal stated that he or she would change a student's placement "when all parties agree and it's truly in the best interest of the child."

Some principals (n = 50/306, 16.3%) also referred to conflicts between students as a legitimate reason to honor parental requests for placement in separate classrooms or even change a placement during the school year. One principal described a rare instance where he or she might consider a new placement necessary, namely, "if there is a bullying issue in the classroom or conflict with another student that [could not] be resolved with regular inventions."

Discussion

In terms of random assignment, when examined in this context, researchers found that many principals (n = 218/321, 67.9%) strongly opposed random placement. While a quarter (n = 81/321, 25.2%) of respondents acknowledged that random methods may have some benefits, they also noted that random placement practices contradict their own educational philosophies.

Some principals emphasized the importance of "balanced" classrooms that can be constructed more purposefully. One explained that "balanced classes relative to academic ability, behavior, special needs, gender, etc. create an even playing field for the teachers relative to achievement." Another agreed, noting that "I try to keep all classes balanced between gender, Title 1 students, special education, behavior, etc." Another described the potentially negative impact of random placement, suggesting that "student placement can make or break a student's learning so I place all students, even moving ones after the year has started." Inversely, however, other principals reported valuing more homogenous classrooms, when they felt such (e.g., cluster grouping) practices were in students' best interests. While bias seems to exist more often when more homogenous groups of students are placed into classrooms, this could present a biasing issue here (Baker et al., 2010; Capitol Hill Briefing, 2011; Goldhaber et al., 2012; Kupermintz, 2003; McCaffrey et al., 2004; Newton et al., 2010; Rothstein & Mathis, 2013; Stacy et al., 2012).

Principals insisted that what they deemed as purposeful placement ensured classrooms that were in students' best interests, encouraged teacher success, and ultimately promoted student learning and achievement. Almost half of respondents (n = 154/321, 48.0%) also believed that random assignment methods, if ever mandated or required for experimental purposes, would prove impractical (n = 57/321, 17.8%) and even detrimental (n = 77/321, 24.0%). This may be what occurred with the aforementioned Bill and Melinda Gates Foundation's (2013) MET study, where the majority of participating schools and teachers reneged on their agreements to safeguard and adhere to the study's randomized design (see also Rothstein & Mathis, 2013).

In addition, some principals (n = 57/332, 17.2%) specifically cautioned against the imbalanced classes likely to result from random assignment. One respondent noted that a "luck of the draw" approach would only "build inequity." One principal summarized his or her commitment to individualized, albeit time-consuming procedures, noting: "I would much rather take the time and find a good match between a student and teacher . . . it is very important that we have a suitable match that is a win-win for everyone." Another principal explained:

> You get what effort you put in. That is if you just shuffle the deck and assign them, you are in for a big mess. Put in the hard work up front and receive the benefits in the end. Plus, one teacher may be strong in reading instruction and that is what certain students need. Why would you not give them that teacher?

Echoing this sentiment, another added:

> I think that random assignment to a classroom is unthinkable. This day in age when we have so much information (data) on students,

> we need to use that information to make all decisions in order to offer the best education possible for each student.

Another expressed his or her disapproval, stating, "I prefer careful, thoughtful, and intentional placement [of students] to random. I've never considered using random placement. These are children, human beings." Another respondent explained that "anything done randomly will get random results. If assignment of students is done strategically with a goal in mind (student success) then there is a higher likelihood of meeting that goal."

With a majority of respondents (n = 218/321, 67.9%) rejecting the practice of random placement, it is evident that even if principals value fairness, equity, and justice, which most if not all of them certainly do, random assignment will probably never be the professionals' practice of choice (see also Burns & Mason, 1995). This unquestionably has implications, particularly for researchers who argue (in many ways correctly) that value-added analyses will probably never be done well without random assignment practices in place (Corcoran, 2010; Glazerman & Potamites, 2011; Reardon & Raudenbush, 2009; Rothstein, 2009, 2010).

In terms of bias, the fundamental question here was whether the nonrandom student assignment practices discovered in this study might logically lead to biased VAM estimates, if the nonrandom student sorting practices went beyond that which is typically controlled for in most VAM models (e.g., academic achievement and prior demonstrated abilities, special education status, ELL status, gender, giftedness). Here, researchers found that while that which is typically controlled for in most VAM models is typically valued here when principals and teachers assign students to classrooms, using these variables to sort students into classrooms is done in highly idiosyncratic and personal ways. It is not that principals, for example, use students' prior academic achievement records and systematically sort students into classrooms. Rather, principals reported considering a wide variety of student factors and variables, including but not limited to the variables for which VAM researchers typically control, when working alongside teachers to make subjective and highly individualized student placement decisions.

Very few principals, for example, identified students' racial, ethnic, and socioeconomic backgrounds as factors they considered in the placement process, which might make controls for these variables less necessary than assumed. Yet, many principals identified students' perceived behavioral needs as major factors that were consistently considered, although often behavior records that might effectively capture students' behavior, discipline, suspension records, and other self-handicapping behaviors are not often available to serve as statistical controls. This goes without mentioning how inconsistent these records might be across varying classrooms, schools, and districts as per their variable student discipline policies and procedures.

Related, some principals reported using students' prior grades to make placement decisions. But whether students' grades can be effectively captured using students' prior test scores, mainly given the lower than expected correlations between grades and test scores often caused by grading variation across classrooms, schools, and districts (Ricketts, 2010; Willingham, Pollack, & Lewis, 2000), might also cause concerns about bias in this context.

Principals also frequently identified students' learning styles, personalities, and interpersonal interactions with other students and teachers as significantly critical factors that influence the student placement process. Researchers are unsure what variables might be able to effectively capture any of these considerations or the extent to which without controlling for learning styles, personalities, and prior interpersonal interactions might cause bias. How might methodologists, for example, control for when students are placed with particular teachers "that [a prior teacher felt] the child[ren] would be most successful with?" How might methodologists also take into consideration the more than one-third of principals who noted that they honored more than 80% of parents' requests regarding what they wanted for their children?

Conclusions

It may be our only option, if we are to move forward in this area, to get value-added models as good as we can, or "good enough." This is what those who support the further use of VAMs continue to argue and accept; specifically that VAMs, even in their current yet faulty forms, are already "good enough" to be used for stronger accountability policies and consequential decision-making purposes (Glazerman et al., 2011; Harris, 2011). The "it's not perfect, but it's the best we have logic pervades pro-VAM arguments . . . even in the absence of a discussion of the implications of its admitted faults" (Gabriel & Allington, 2011, p. 7).

At best, perhaps, we might supplement these with other data that also capture teacher effectiveness in line with the educational measurement standards of the profession (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2000). However, this approach is also manifesting itself as substantially more difficult, again, than many would assume. The correlations being demonstrated among the multiple measures in such systems (e.g., that include value-added estimates, teacher observational scores, student surveys of teacher quality, etc.) are unacceptably low (e.g., $r \leq 0.50$; see e.g., Bill and Melinda Gates Foundation, 2013; Kersting et al., 2013), especially if VAMs are to be used for consequential purposes. This is also adding to the legitimate concerns about the validity of all of the limited measures being used for increased accountability purposes, not just the "more scientific" VAMs of interest here.

It is argued herein that the purposeful (nonrandom) assignment of students into classrooms biases value-added estimates and their valid interpretations. Researchers conducted this study in order to understand how often random assignment practices are used to place students into classrooms and to determine how and whether the controls used by value-added researchers to eliminate bias might be sufficient given what indeed occurs in practice.

Researchers found that both teachers and parents play a prodigious role in the student placement process, in almost 9 out of 10 schools. They provide both appreciated and sometimes unwelcome insights, regarding what they perceive to be the best learning environments for their students or children, respectively. Their added insights typically revolve around students' behaviors, learning styles, personalities, and interactions with their peers, prior teachers, and general teacher types (e.g., teachers who manage their classrooms in perceptibly better ways). These things are not typically controlled for across current VAMs. These factors serve as legitimate reasons for class changes during the school year as well, although whether this too could be captured is tentative at best. Otherwise, namely, prior academic achievement, special education needs, giftedness, and gender heavily influence placement decisions. These are variables for which most current VAMs account or control, presumably effectively.

Also of importance was that principal respondents were greatly opposed to using random student assignment methods in lieu of placement practices based on human judgment—practices they collectively agreed were in the students' best interests. Random assignment, even if necessary to produce unbiased VAM-based estimates, was deemed highly nonsensical and impractical. Although some principals acknowledged the potential benefits of randomized methods of assignment, many described random placements as unreasonable and even detrimental to student learning and teacher success. Overall, principals saw random assignment as counterproductive to the students' best interests, although they never considered the inequitable learning environments that might result, given what they themselves perceived as their more appropriate yet highly individualized student placement procedures. Related, despite the complexities of the student placement process, researchers found that assignment methods are rarely discussed in principals' administrative coursework or other professional development training or in district policy manuals. This too has implications in this context as well.

Further research is warranted to help us determine how student placement decisions bias value-added estimates and perhaps how varying practices impact or bias estimates in different ways. But for now, given the widespread use of the nonrandom methods illustrated in this study, value-added researchers, policymakers, and educators (particularly those whose effectiveness is being measured) might more carefully consider the implications

of their placement decisions. They might more carefully consider, as well, the validity of the inferences they make using such potentially biased estimates, since student assignment practices will likely continue to distort the value-added estimates now so widely being adopted and used.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2000). *AERA position statement on high-stakes testing in pre-K–12 education*. Retrieved from http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publication/bp278/

Ballou, D. (2002). Sizing up test scores. *Education Next*. Retrieved from education-next.org/files/ednext20022_10.pdf

Ballou, D. (2012). Review of The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-long-term-impacts

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

Bill and Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

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

Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/due-diligence

Burns, R. B., & Mason, D. A. (1995). Organizational constraints on the formation of elementary school classes. *American Journal of Education*, *103*(2), 185–212. doi:10.1086/444096

Capitol Hill Briefing. (2011). *Getting teacher evaluation right: A challenge for policy makers*. Retrieved from http://www.aera.net/Default.aspx?id=12856

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be?* Providence, RI: Annenberg Institute for School Reform.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Dills, A. K., & Mulholland, S. E. (2010). A comparative look at private and public schools' class size determinants. *Education Economics*, *18*(4), 435–454. doi:10.1080/09645290903546397

Dunn, M. C., Kadane, J. B., & Garrow, J. R. (2003). Comparing harm done by mobility and class absence: Missing students and missing data. *Journal of Educational and Behavioral Statistics*, *28*(3), 269–288. doi:10.3102/10769986028003269

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). *Selecting growth measures for school and teacher evaluations*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Ercikan, K., & Wolff-Michael, R. (Eds.) (2009). *Generalizing from educational research: Beyond qualitative and quantitative polarization*. New York, NY: Routledge.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119–161). New York, NY: Macmillan.

Gabriel, R., & Allington, R. (2011). *Teacher effectiveness research and the spectacle of effectiveness policy*. Paper Presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

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

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1*(1), 2–16.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems*. Retrieved from www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx

Glazerman, S. M., & Potamites, L. (2011). *False performance gains: A critique of successive cohort indicators*. Princeton, NJ: Mathematica Policy Research. Retrieved from www.mathematica-mpr.com/publications/pdfs/.../False_Perf.pdf

Goldhaber, D., Gabele, B., & Walch, J. (2012). *Does the model matter? Exploring the relationship between different student achievement-based teacher assessments*. Seattle, WA: Center for Education Data and Research (CEDR). Retrieved from https://appam.confex.com/appam/2012/webprogram/Paper2264.html

Goldhaber, D., & Theobald, R. (2012). *Do different value-added models tell us the same things?* Palo Alto, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P., & Wooldridge, J. M. (2012). *An evaluation of Empirical Bayes' estimation of value-added teacher performance measures*. East Lansing, MI: Education Policy Center. Retrieved from www.aefpweb.org/sites/default/files/webform/empirical_bayes_20120301_AEFP.pdf

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher education performance be trusted?* East Lansing, MI: Education Policy Center. Retrieved from http://education.msu.edu/epc/library/documents/WP18Guarino-Reckase-Wooldridge-2012-Can-Value-Added-Measures-of-Teacher-Performance-Be-T_000.pdf

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An evaluation of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*, 319–350. doi:10.1162/edfp.2009.4.4.319

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

Harris, D. N., & Anderson, A. (2013). *Does value-added work better in elementary than in secondary grades?* Palo Alto, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/grades/

Hermann, M., Walsh, E., Isenberg, E., & Resch, A. (2013). *Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/publications/PDFs/education/value-added_shrinkage_wp.pdf

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

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

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Cambridge, MA: The National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Kersting, N. B., Chen, M., & Stigler, J. W. (2013). Value-added added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives*, *21*(7), 1–39. Retrieved from http://epaa.asu.edu/ojs/article/view/1167

Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function*. Nashville, TN: National Center on Performance Initiatives. Retrieved from http://ideas.repec.org/p/umc/wpaper/0708.html

Koedel, C., & Betts, J. R. (2010). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy 6*(1), 18–42. doi:10.1162/EDFP_a_00027

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation & Policy Analysis*, *25*(3), 287–298. doi:10.3102/01623737025003287

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Palo Alto, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/level-playing-field/

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. doi:10.3102/10769986029001067

Meyer, R. H., & Dokumaci, E. (2010). *Value-added models and the next generation of assessments*. Princeton, NJ: Center for K–12 Assessment & Performance Management. Retrieved from www.k12center.org/rsc/pdf/MeyerDokumaciPresenterSession4.pdf

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models

and contexts. *Educational Policy Analysis Archives*, *18*(23). Retrieved from http://epaa.asu.edu/ojs/article/view/810

Office of the Auditor General, State of Arizona. (2007). *Baseline study of Arizona's English Language Learner programs and data*. Retrieved from http://www.auditorgen.state.az.us/Reports/School_Districts/Statewide/2008_April/ELL_Baseline_Report.pdf

Player, D. (2010). Nonmonetary compensation in the public teacher labor market. *Education Finance and Policy*, *5*(1), 82–103. doi:10.1162/edfp.2009.5.1.5105

Praisner, C. (2003). Attitudes of elementary school principals toward the inclusion of students with disabilities. *Exceptional Children*, *69*(2), 135–145.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service (ETS). Retrieved from www.ets.org/Media/Education_Topics/pdf/angoff9.pdf

Raudenbush, S. W., & Jean, M. (2012). *How should educators interpret value-added scores?* Palo Alto, CA: Carnegie Knowledge Network. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519. doi:10.1162/edfp.2009.4.4.492

Ricketts, C. R. (2010). *End of course grades and standardized test scores: Are grades predictive of student achievement?* (Doctoral dissertation). Available from ProQuest LLC. Retrieved from http://search.proquest.com/docview/755303452

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. doi:10.1111/j.1468-0262.2005.00584.x

Rivkin, S. G., & Ishii, J. (2008). *Impediments to the estimation of teacher value added*. Paper presented at the National Conference on Value-Added Modeling. Sponsored by the Wisconsin Center for Education Research (WCER), Madison, WI.

Rothstein, J. (2009, January 11). *Student sorting and bias in value-added estimation: Selection on observables and unobservables*. Cambridge, MA: The National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214. doi:10.1162/qjec.2010.125.1.175

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116. doi:10.3102/10769986029001103

Sanders, W. L. (1998). Value-added assessment. *The School Administrator*, *55*(11), 24–27.

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247–256.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. Cary, NC: SAS Institute Inc. Retrieved from www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, *95*(2), 122–140. doi:10.1177/0192636511410052

Stacy, B., Guarino, C., Recklase, M., & Wooldridge, J. (2012). *Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?* East Lansing, MI: Education Policy Center. Retrieved from https://appam.confex.com/appam/2012/webprogram/Paper3327.html

Stake, R. E., & Trumbull, D. (1982). Naturalistic generalizations. *Review Journal of Philosophy and Social Science*, *7*, 1–12.

State of Arizona Department of Education. (2010–2011). *State report card*. Retrieved from http://www.azed.gov/research-evaluation/files/2012/04/2011statereportcard.pdf

Strauss, A. L., & Corbin, J. (1995). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, *13*(1), 12–28.

Thompson, B. (2000). *The APA Task Force on Statistical Inference Report as a framework for teaching and evaluating students' understandings of study validity*. Paper presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA.

U.S. Census Bureau. (2011). *State and county quick-facts*. Retrieved from http://quickfacts.census.gov/qfd/states/04000.html

U.S. Department of Education, National Center for Education Statistics (NCES), Common Core of Data (CCD). (1999–2000 and 2009–2010). *State nonfiscal survey of public elementary/secondary education (Table 44)*. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_044.asp

U.S. Department of Education, National Center for Education Statistics (NCES), Common Core of Data (CCD). (2009–2010a). *Public elementary/secondary school universe survey (Table 45)*. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_045.asp

U.S. Department of Education, National Center for Education Statistics (NCES), Common Core of Data (CCD). (2009–2010b). *Public elementary/secondary school universe survey (Table 48)*. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_048.asp

U.S. Department of Education, National Center for Education Statistics (NCES), Common Core of Data (CCD). (2009–2010c). *Public elementary/secondary school universe survey (Table 104)*. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_104.asp

U.S. Department of Education, National Center for Education Statistics (NCES), Common Core of Data (CCD). (2010–2011). *Public elementary/secondary school universe survey (Table 2)*. Retrieved from http://nces.ed.gov/pubs2012/pesschools10/tables/table_02.asp

U.S. Department of Education, National Center for Education Statistics (NCES), Schools and Staffing Survey. (2007–2008). *Characteristics of public, private, and Bureau of Indian Education elementary and secondary school principals in the United States (Table 7)*. Retrieved from http://nces.ed.gov/pubs2009/2009323/tables/sass0708_2009323_p12n_07.asp

U.S. Department of Education, Office for Civil Rights. (2004, 2006). *Civil rights data collection (Table 50)*. Retrieved from http://nces.ed.gov/programs/digest/d11/tables/dt11_050.asp

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. doi:10.1037//0003-066X.54.8.594

Willingham, W. W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences*. Princeton, NJ: Education Testing Service (ETS). Retrieved from http://www.ets.org/Media/Research/pdf/RR-00-15-Willingham.pdf

Word, E., Johnston, J., Bain, H. P., Fulton, D. B., Zaharias, J. B., Achilles, C. M., . . . Breda, C. (1990). *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) project*. Nashville, TN: Tennessee State Department of Education. Retrieved from http://d64.e2services.net/class/STARsummary.pdf

Wright, P., Horn, S., & Sanders, W. L. (1997). Teachers and classroom heterogeneity: Their effects on educational outcomes. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Manuscript received November 1, 2012
Final revision received July 25, 2013
Accepted September 13, 2013