A Reanalysis of the Effects of Teacher Replacement Using Value-Added Modeling

STUART S. YEH

University of Minnesota

Background: *In principle, value-added modeling (VAM) might be justified if it can be shown to be a more reliable indicator of teacher quality than existing indicators for existing low-stakes decisions that are already being made, such as the award of small merit bonuses. However, a growing number of researchers now advocate the use of VAM to identify and replace large numbers of low-performing teachers. There is a need to evaluate these proposals because the active termination of large numbers of teachers based on VAM requires a much higher standard of reliability and validity. Furthermore, these proposals must be evaluated to determine if they are cost-effective compared to alternative proposals for raising student achievement. While VAM might be justified as a replacement for existing indicators (for existing decisions regarding merit compensation), it might not meet the higher standard of reliability and validity required for large-scale teacher termination, and it may not be the most cost-effective approach for raising student achievement. If society devotes its resources to approaches that are not cost-effective, the increase in achievement per dollar of resources expended will remain low, inhibiting reduction of the achievement gap.*

Objective: *This article reviews literature regarding the reliability and validity of VAM, then focuses on an evaluation of a proposal by Chetty, Friedman, and Rockoff to use VAM to identify and replace the lowest-performing 5% of teachers with average teachers. Chetty et al. estimate that implementation of this proposal would increase the achievement and lifetime earnings of students. The results appear likely to accelerate the adoption of VAM by school districts nationwide. The objective of the current article is to evaluate the Chetty et al. proposal and the strategy of raising student achievement by using VAM to identify and replace low-performing teachers.*

Method: *This article analyzes the assumptions of the Chetty et al. study and the assumptions of similar VAM-based proposals to raise student achievement. This analysis establishes a basis for evaluating the Chetty et al. proposal and, in general, a basis for evaluating all VAM-based policies to raise achievement.*

Teachers College Record Volume 115, 120305, December 2013, 35 pages

0161-4681

Conclusion: *VAM is not reliable or valid, and VAM-based polices are not cost-effective for the purpose of raising student achievement and increasing earnings by terminating large numbers of low-performing teachers.*

INTRODUCTION

Value-added modeling (VAM) may be defined as the use of statistical methods for the purpose of isolating the "value-added" contribution of individual teachers to student achievement (Braun, Chudowsky, & Koenig, 2010). Numerous researchers advocate the use of value-added performance information to make decisions about hiring, firing, rewarding, or promoting teachers (Glazerman, Goldhaber, Loeb, Staiger, & Whitehurst, 2010; Gordon, Kane, & Staiger, 2006; Hanushek, 2009a, 2010; Hess, 2010; Staiger & Rockoff, 2010). School districts across the nation are adopting VAM and many districts are using VAM for high-stakes decisions (Dillon, 2010; The Center for Greater Philadelphia, 2004). Tennessee has incorporated VAM into teacher evaluations since 1998 (Tennessee Department of Education, 2011). District of Columbia Public Schools uses VAM in teacher evaluations (District of Columbia Public Schools, 2012), along with 29 District of Columbia public charter schools (Turque, 2012). Pittsburgh Public Schools uses VAM in teacher evaluations (Johnson, Lipscomb, Gill, Booker, & Bruch, 2012). The states of New York, Louisiana, and Colorado and the Los Angeles school district plan to incorporate VAM into teacher evaluations (Colorado State Council for Educator Effectiveness, 2011; Louisiana Department of Education, 2011; New York State Department of Education, 2011; Watanabe, 2011). Federal policy endorses this approach by directing federal funds to states that adopt the approach (Dillon, 2010; U.S. Department of Education, 2012).

Given the rapid adoption of VAM in districts across the nation and its apparent endorsement by researchers and policymakers, there is a need to evaluate the effectiveness of policies based on VAM. One study in particular (Chetty, Friedman, & Rockoff, 2011) has received a great amount of attention (Lowrey, 2012). Chetty et al. (2011) suggest that the use of VAM to identify and replace the lowest-performing 5% of teachers with average teachers would increase student achievement and would translate into sizable gains in the lifetime earnings of their students: "The total undiscounted earnings gains from this policy are $52,000 per child and more than $1.4 million for the average classroom" (Chetty et al., 2011, p. 5).[1] These startling figures have been cited to justify the use of

VAM and appear likely to accelerate the adoption of VAM by school districts nation-wide (Kristof, 2012a, 2012b).

If policies based on VAM are indeed as effective as indicated by the Chetty et al. (2011) study, then national implementation would appear to be warranted. However, there is a need to critically examine the study's assumptions. If those assumptions are incorrect, then the conclusions may be incorrect, and national implementation may serve to divert scarce resources from improvement strategies that are more effective. While it may be the case that relaxing the assumptions does not alter Chetty et al.'s conclusions, it is incumbent upon researchers to demonstrate that this is true. Ballou (2012) noted that the Chetty et al. study applied few quasi-experimental tests to rule out the possibility that high value-added teachers had been systematically assigned students whose increased earnings are attributable to factors other than differences in teacher value-added. In the absence of these tests, Ballou concluded that it is not appropriate to attribute increased earnings to measured differences in teacher value-added. The present article goes beyond Ballou's critique to identify multiple assumptions that are implicit when VAM is used to identify and replace low-performing teachers.

An important distinction is whether VAM is only used to replace existing indicators of teacher quality, such as principal judgments, for existing decisions (regarding merit pay, for example) that are already being made or whether it is used to justify a large expansion of teacher termination and replacement, as in the case of the Chetty et al. (2011) proposal. It may be the case that VAM is a better predictor than other factors that are currently used to make decisions about pay, promotion, or hiring and, therefore, can be justified as a better substitute for those other predictors. However, the controversial aspect of VAM is its expanded use as an explicit strategy for terminating teachers who would not otherwise be terminated, in an effort to improve student achievement. What is missing from this discussion are analyses to determine whether this strategy is a cost-effective use of society's scarce resources compared to alternative strategies for raising student achievement. VAM might be justified as the best predictor of teacher quality for decisions that are already being made using less reliable predictors but may not be justifiable for policies that involve vast expansion of teacher termination and replacement, unless this strategy has been shown to be the most cost-effective approach for raising student achievement.

Section 1 of this article reviews literature regarding the reliability and validity of VAM. Section 2 analyzes several key assumptions underlying the Chetty et al. (2011) study and suggests that these assumptions are common to studies that evaluate the effectiveness of policies based on

VAM. Section 2 includes a range of cost-effectiveness and benefit-cost analyses of the Chetty et al. proposal. Section 3 concludes that VAM is neither reliable nor valid for the purpose of high-stakes decisions regarding teacher hiring and firing, and VAM-based policies are not cost-effective strategies for raising student achievement. In view of the need to consider alternatives, section 3 compares VAM-based policies to rapid performance feedback (RPF), which appears to be far more cost-effective and suggests an alternative way of thinking about strategies for improving student achievement.

1. RELIABILITY AND VALIDITY ISSUES

Interest in VAM was stimulated by Sanders and Rivers (1996), who used statistical methods to isolate the contribution of individual teachers to student achievement two years into the future. This suggested that teachers have persistent effects on their students' achievement and that the accumulation of these effects could be substantial. The following year, Sanders and his colleagues published an article asserting that teachers are the most important factor influencing student achievement (Wright, Horn, & Sanders, 1997). Interest in VAM grew as subsequent studies indicated that the contribution of teachers to student achievement is large, and value-added estimates of teachers' contributions predict their students' measured achievement (Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002; Staiger & Rockoff, 2010). Teacher ratings based on VAM are moderately correlated with ratings obtained from portfolio evidence and classroom observations conducted by trained evaluators (Hill, Kapitula, & Umland, 2011; see also Jacob & Lefgren, 2008; Milanowski, 2004; Schacter & Thum, 2004). The evidence that teachers have significant effects on student achievement led many researchers to advocate the use of VAM to identify and replace low-performing teachers (Gordon et al., 2006; Hanushek, 2009a, 2010; Staiger & Rockoff, 2010). Using Monte Carlo simulations, Staiger and Rockoff (2010) asserted that "80 percent of teachers should be dismissed after their first year" based on VAM estimates of their effectiveness (p. 108). Advocates of using VAM for high-stakes decisions regarding teacher hiring and firing argue that concerns about falsely identifying low-performing teachers can be addressed by using multiple years of data to estimate each teacher's ranking and that an excessive concern with false identifications serves the interests of teachers, rather than their students (Glazerman, Loeb, et al., 2010).

Any VAM-based policy to identify and replace low-performing teachers, however, requires the strong assumption that specific teachers *cause* the observed gains or losses in their students' achievement (Braun et al., 2010). The critical assumption is that any differences among classes, schools, or programs that are not captured by the predictor variables used in the VAM model are captured by the student fixed-effect components (Braun et al., 2010). However, using data from North Carolina, Rothstein (2009, 2010) found that the estimated effect for fifth-grade teachers predicts their students' prior performances. Since it is impossible for fifth-grade teachers to cause performance that occurred prior to the fifth grade, this result implies there is nonrandom selection of students into teacher classrooms that is not controlled through the inclusion of time-invariant student characteristics. Therefore, the central assumption underlying VAM appears to be invalid (Braun et al., 2010). Rothstein concluded: "Results indicate that even the best feasible value-added models may be substantially biased" (Rothstein, 2009, p. 537). This surprising result suggests that the use of VAM to identify and terminate low-performing teachers is not warranted. When teachers are assigned students who achieved high gains in performance the previous year, existing VAM models erroneously subtract a portion of the gain that is properly attributed to these teachers, making them look like bad teachers (Rothstein, 2009). This problem may be exacerbated if VAM is used to identify and terminate teachers because the high stakes may cause teachers to lobby principals for students who are predicted to post large gains in the coming year, and principals may be tempted to use their control over classroom assignments to reward favored teachers (Koedel & Betts, 2011; Rothstein, 2009).

Using data from San Diego, Koedel and Betts (2011) corroborated Rothstein's (2009) primary finding, demonstrating that the effect is not unique to North Carolina. However, Koedel and Betts also found that sorting bias can be almost completely mitigated when a complex value-added model is used that restricts the analysis to teachers for whom at least three contiguous cohorts of student test scores are available. A major difficulty, however, is that it would not be uncommon for data to be missing in a way that would prevent the use of this technique with large numbers of teachers (Newton, Darling-Hammond, Haertel, & Thomas, 2010). Not only would it be necessary for teachers to have three contiguous cohorts of student test scores, but most VAM models are restricted to complete cases of data, which is only appropriate if the missing data are missing completely at random (Rubin, Stuart, & Zanutto, 2004). This assumption is inappropriate because systematic factors influence missing school data. For example, students who move may be more likely to be students who perform at lower levels.

Ishii and Rivkin (2009) identified specific parent and school influences on student assignment to classrooms that may systematically bias VAM estimates even when the models incorporate student fixed effects. Highly-educated parents are more likely to request that their children be assigned to particular teachers. Highly-educated parents may also hire tutors during time periods when they perceive that their children's teachers are inadequate. Also, teachers tend to prefer classrooms with higher-achieving students, and principals might assign high-performing teachers to classrooms with high-achieving students as an incentive for the high-performing teachers to remain at a school. Not all of these influences could be controlled using student fixed effects because the purposeful nature of these choices almost certainly introduces correlations among teacher quality and family/student characteristics (Ishii & Rivkin, 2009).

Employing the same tests used by Rothstein (2009, 2010), Briggs and Domingue (2011) analyzed the VAM model developed by the RAND Corporation and used by the Los Angeles Unified School District to rank teachers. Briggs and Domingue found that estimates produced by the model are significantly biased and teacher rankings are highly dependent on the specification of the model. An alternative specification controlling for a longer history of each student's performance, peer influences, and school-level factors produced different teacher ratings: In reading, 53.6% of teachers did not retain the same effectiveness rating under both specifications; in math, 39.2% of teachers did not retain the same effectiveness rating. This suggests that teacher ratings using VAM are highly sensitive to details regarding the model's implementation.

Ballou, Sanders, and Wright (2004) point out that the inclusion of socioeconomic status (SES) in an effort to control for differences in family/student characteristics would bias any estimate of teacher effectiveness toward zero because of the likely correlation between SES and teacher quality. For this reason, the Education Value-Added Assessment System (EVAAS), a popular variant of VAM, omits student covariates including SES. However, McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) found that this would likely confound estimated teacher effects, and teacher rankings based on these effects, when different schools serve distinctly different student populations. Ballou et al. point out that EVAAS, which uses each student's test score history to substitute for SES and demographic variables, is not vulnerable to missing SES and demographic data and, in Tennessee, produced teacher rankings that were comparable to rankings when SES and demographic variables were included. However, no systematic study has examined EVAAS rates of false positive and false negative teacher classifications (Kupermintz, 2003).

Another problem with VAM is that it does not appear possible to separate teacher and school effects using currently available accountability data (Raudenbush, 2004). Separating these effects would only be possible if each teacher regularly taught at multiple schools where the accountability systems were consistent and the data were available across schools. Currently, however, when VAM is used to estimate individual teacher effects and to rank teachers, these estimates are contaminated by effects that are properly attributed to schools, not teachers. Furthermore, there is no obvious solution to this problem.

A largely ignored problem is that *true* teacher performance, contrary to the main assumption underlying current VAM models, varies over time (Goldhaber & Hansen, 2012). These models assume that each teacher exhibits an underlying trend in performance that can be detected given a sufficient amount of data. The question of stability is not a question about whether *average* teacher performance rises, declines, or remains flat over time. The issue that concerns critics of VAM is whether *individual* teacher performance fluctuates over time in a way that invalidates inferences that an individual teacher is "low-" or "high-" performing. This distinction is crucial because VAM is increasingly being applied such that individual teachers who are identified as low-performing are to be terminated. From the perspective of individual teachers, it is inappropriate and invalid to fire a teacher whose performance is low this year but high the next year, and it is inappropriate to retain a teacher whose performance is high this year but low next year. Even if average teacher performance remains stable over time, individual teacher performance may fluctuate wildly from year to year.

While previous studies examined the intertemporal stability of value-added teacher rankings over one-year periods and found that reliability is inadequate for high-stakes decisions, researchers tended to assume that this instability was primarily a function of measurement error and sought ways to reduce this error (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009). However, this hypothesis was rejected by Goldhaber and Hansen (2012), who investigated the stability of teacher performance in North Carolina using data spanning 10 years and found that much of a teacher's true performance varies over time due to unobservable factors such as effort, motivation, and class chemistry that are not easily captured through VAM. This invalidates the assumption of stable teacher performance that is embedded in Hanushek's (2009b) and Gordon et al.'s (2006) VAM-based policy proposals, as well as VAM models specified by McCaffrey et al. (2009) and Staiger and Rockoff (2010) (see Goldhaber & Hansen, 2012, p. 15). The implication is that standard estimates of impact when using VAM to identify and replace low-performing teachers are significantly inflated (see Goldhaber & Hansen, 2012, p. 31).

Another problem arises when, for example, a pretest score measures pre-algebra but the posttest score measures geometry skills or when a teacher emphasizes pre-algebra but not geometry. Improvements in learning may not be captured by the assessment. A mismatch between instruction and assessment would tend to invalidate VAM-based teacher rankings (Reckase, 2004). VAM requires the use of vertically-scaled achievement data that spans wide grade, developmental, and content ranges. However, the shift in constructs that are measured from grade to grade introduces remarkable distortions: Effective teachers may be identified as ineffective and vice-versa, and effects contributed by prior teachers may be erroneously attributed to later teachers (Martineau, 2006). Martineau writes: "With current technology, there are no vertical score scales that can be validly used in high-stakes analyses for estimating value added to student growth in either grade-specific or student-tailored construct mixes . . . . A serious (but reasonable) implication of this study is to all but eliminate the high-stakes use of value-added accountability systems based on vertically scaled student achievement data" (2006, pp. 57-58). Even when instruction and assessment are matched, differences in the particular achievement tests that are used produce substantially different answers about individual teacher performance and do not rank teachers consistently (Papay, 2011).

2. ASSUMPTIONS

The preceding review of literature suggests numerous reasons for caution in using the results of any VAM model to identify and replace low-performing teachers. These concerns are magnified when VAM is used, as it is used in the Chetty et al. (2011) study, to make assertions about the long-term economic benefits to students who are taught by teachers identified as "high-performing" teachers according to the VAM analysis. The analysis presented in this article suggests that the findings of the Chetty et al. study depend on numerous assumptions that may be questioned. Significantly, these assumptions are common to studies that predict positive benefits of policies based on VAM. Therefore, the analysis presented here has implications for VAM-based policies in general whenever they are used to make predictions about the long-term benefits of identifying and replacing low-performing teachers.

2.1 FIXED TEACHER QUALITY?

A key assumption of the Chetty et al. (2011) analysis is that true teacher quality is fixed over time: "The model for scores . . . . assumes that teacher quality μ_j is fixed over time . . . . This rules out the possibility that teacher quality fluctuates across years" (p. 7). In other words, the Chetty et al. analysis assumes that a high-quality teacher this year will remain a high-quality teacher next year; a low-quality teacher this year will remain a low-quality teacher next year. Later in the article, however, the authors conclude based on their data that "teacher value-added is not in fact a time-invariant characteristic" (p. 25). While the authors' analysis assumed that teacher quality is fixed over time, their own data suggest that teacher quality, as measured by teacher value-added "is not in fact" time-invariant, consistent with the results reported by Goldhaber and Hansen (2012). If this assumption is not valid, the conclusions of the analysis are not likely to be valid.

The intertemporal reliability of value-added teacher rankings was investigated by Aaronson et al. (2007), Ballou (2005), Koedel and Betts (2007), and McCaffrey et al. (2009). In each study, VAM was used to rank teacher performance from high to low. In each study, a majority of teachers who ranked in the lowest quartile or lowest quintile shifted out of that quartile (or quintile) the following year (see Tables 1 and 2). Furthermore, a majority of teachers who ranked in the highest quartile or quintile shifted out of that quartile (or quintile) the following year (see Tables 1 and 2).

Table 1. Instability of Value-Added Teacher Rankings in Chicago and Tennessee

	Teacher Rankings	
Locale	Bottom 25% in Year t; Top 75% in Year t+1	Top 25% in Year t; Bottom 75% in Year t+1
Chicago, IL	67%	59%
Tennessee	60%	52%

Notes. Chicago data are from Aaronson et al. (2007, Table 7) for high school math teachers, with controls for student, peer, and neighborhood covariates. Tennessee data are from Ballou (2005, Figure 5b) for math teachers in grades 3–8 in a single large district.

Table 2. Instability of Value-Added Teacher Rankings in San Diego and 5 Florida Counties

	Teacher Rankings	
Locale	Bottom 20% in Year t; Top 80% in Year t+1	Top 20% in Year t; Bottom 80% in Year t+1
San Diego, CA	65%	71%
Dade County, FL	70%	67%
Duval County, FL	67%	61%
Hillsborough County, FL	67%	67%
Orange County, FL	59%	65%
Palm Beach County, FL	69%	68%

Notes. San Diego data are from Koedel and Betts (2007, Table 9) based on elementary school math teachers, with controls for student and school fixed effects. Data for Florida counties are from McCaffrey et al. (2009, Table 4) based on elementary school math teachers with 15 or more students per year, with controls for student fixed effects.

What this means is that value-added teacher rankings are insufficiently reliable for the purpose of high-stakes decisions regarding hiring and firing. High-stakes decisions are clearly unwarranted if this volatility in the rankings is due to unmeasured variables or random measurement error. However, even in the unlikely event that there are no unmeasured variables and measurement error is zero, implying that all volatility is due to true variation in teacher performance, it would not be appropriate to hire or fire based on the ranking in a given year (designated "year *t*"). In over half of all instances, performance would have either improved or declined the following year (designated "year *t*+1") by such an extent as to invalidate the year *t* ranking. If VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 1 and 2 indicate that this decision is incorrect, according to the year *t*+1 teacher rankings, between 59 and 70 % of the time. If VAM-based culling is less reliable than flipping a coin, as these results suggest, then productive teachers would be culled more frequently than unproductive bottom quartile (or bottom quintile) teachers.[2]

In the case of value-added rankings, *it is inappropriate* to infer that a teacher should be hired or fired based on the rankings from any given year. Since this inference would be inappropriate, the results of value-added teacher rankings are not valid for the purpose of high-stakes decisions regarding hiring and firing.[3] In short, VAM lacks validity for the purpose of high-stakes decisions regarding individual teachers.

While some researchers suggest averaging two or more years of rankings to improve reliability, averaging may introduce significant bias—raising the issue of validity once again (McCaffrey et al., 2009). Furthermore, it would not be uncommon for data to be missing in a way that would prevent averaging. For large numbers of teachers, it would be impractical to average teachers' rankings across two or more years (Newton et al., 2010). Regardless, when two years of rankings are used for tenure decisions, intertemporal reliability remains low: In reading, data from North Carolina indicate that 68% of teachers ranked in the bottom quintile shift out of that quintile after tenure (indicated by a weighted average of all post-tenure observations), and 54% of teachers ranked in the top quintile shift out of that quintile post-tenure (Goldhaber & Hansen, 2008). When three years of rankings are used, reliability is even worse: 74% of teachers ranked in the bottom quintile shift out of that quintile post-tenure, and 56% of teachers ranked in the top quintile shift out of that quintile post-tenure (Goldhaber & Hansen, 2008). In math, reliability is somewhat better, but over half of all teachers in the bottom and top quintiles shift out of those quintiles post-tenure (Goldhaber & Hansen, 2008).

These results were confirmed by a second value-added analysis, also using data from North Carolina, which found that more than half of all teachers who ranked in the bottom quintile shifted out of that quintile the following year, regardless of whether one, two, three, four or five years of data were used to predict future performance, regardless of the subject area (math or reading), and regardless of whether a simple or complex Bayes estimator was used to improve predictive accuracy (Lefgren & Sims, 2012).

2.2 ISOLATED TEACHER IMPACT?

Chetty et al. (2011) interpret their results as if the impact of an individual teacher can be isolated: “In this pooled regression, the coefficient estimate β represents the mean impact of having a higher value-added teacher for a single grade between grades 4-8” (p. 36). Chetty et al. interpret the coefficient as follows: “A 1 standard deviation (SD) increase in teacher value-added in a single grade increases earnings at age 28 by $182, 0.9% of mean earnings” (p. 39). This interpretation is the foundation for their statement that replacing a low-quality teacher with a high-quality teacher would result in a large lifetime gain in income for each class of students taught by this teacher (Chetty et al., 2011, p. 48). Later, however, Chetty et al. acknowledge, due to limitations in their analytical method, that it is not valid to interpret β (or the net impact estimate of β) as if the impact of teacher quality has been isolated from the influence of all other inputs (pp. 12, 46). As a consequence, factors other than teacher quality may explain the $182 gain in earnings at age 28. Chetty et al. point out that some of the impact may be due, for example, to the influence of parental social connections that permit children from wealthier families to obtain higher-paying jobs. This influence was not controlled in the Chetty et al. analysis, nor is there an obvious methodological remedy that could be applied by other researchers—suggesting that the problem is not easily corrected. The need to control for social connections is especially important because even a weak influence from connections might explain a small $182 difference in annual earnings.

2.3 CONSEQUENTIAL IMPACT?

Without adequate controls, Chetty et al.’s (2011) estimate of the impact of raising teacher quality by one standard deviation may be questioned. In any case, the estimated impact is quite small. With regard to student achievement, a one-unit increase in teacher quality is associated with a 0.843 standard deviation increase in student test scores (Chetty et al., 2011, Table 4). Since a 0.1 unit increase in teacher quality is equal to a one standard deviation increase in teacher quality, this implies that a one standard deviation increase in teacher quality is associated with a small 0.0843 standard deviation increase in student test scores (Chetty et al., 2011, p. 24).

If VAM is used to replace the lowest 10% of all teachers, any gains in student performance would be limited to 10% of all students. A hypothetical 0.0843 standard deviation gain in performance for 10% of all students would translate, in the aggregate, to an average 0.00843 standard deviation gain for all students, or approximately six days of learning over one academic year.

With regard to earnings, Chetty et al. (2011) estimated that a one standard deviation increase in teacher quality is associated with a 0.9% increase in income at age 28, equal to $182 for a single person. Assuming that this differential persists at every age throughout a person's life, Chetty et al. estimated that the cumulative lifetime gain for a single person would equal $4,600 after discounting the gains at an annual rate of 3% (Chetty et al., 2011, p. 39). Once again, if VAM is used to replace the lowest 10% of all teachers, any gains would be limited to 10% of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $460 per person, averaged across all students.

Chetty et al. (2011) estimated that a larger 2.04 standard deviation increase in teacher quality is associated with a $9,422 cumulative lifetime gain for a single person after discounting the gains at an annual rate of 3% (p. 48). This equals $266,643 for an entire class of 28.3 students. If (as suggested by Chetty et al.) VAM is used to replace the lowest 5% of all teachers, any gains would be limited to 5% of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $471 per person, averaged across all students.

While the preceding analysis suggests that the impact on lifetime earnings averaged over all students would be small, newspaper accounts focused on the claim that the use of VAM to identify and replace the lowest-performing 5% of teachers with average teachers would translate into much larger gains in the lifetime earnings of their students (Kristof, 2012a, 2012b). Indeed, Chetty et al. (2011) state that "The total undiscounted earnings gains from this policy are $52,000 per child and more than $1.4 million for the average classroom" (p. 5). How can this be reconciled with the view that gains are small?

The explanation is that the $52,000 and $1.4 million figures were not discounted to reflect the time value of money. Income received many years in the future is not as valuable as income that is received today. For this reason, economists discount future income streams, effectively reducing the amounts to account for the time value of money. Chetty et al. (2011) reported that after discounting at a 3% annual rate, the lifetime gain of $52,000 per child shrinks to $9,422; the lifetime gain of $1.4 million for an entire classroom of 28.3 students shrinks to $266,643. The smaller amounts are the appropriate amounts to use in any economic analysis of the benefits and costs of VAM-based policies. Once the $9,422 figure is averaged over all students, it shrinks further to $471 per person.

2.4 STABLE QUALITY DIFFERENTIALS?

As noted, Chetty et al. (2011) estimated that substituting an average teacher for a teacher in the bottom 5% of all teachers would result in a lifetime gain, after discounting, equal to $266,643 for a class of 28.3 students taught by that teacher. It may be argued that, regardless of the analysis in section 2.3, a gain of $266,643 remains significant. However, the working assumption is that a teacher in the bottom 5% consistently performs at a level that is 2.04 standard deviations below an average teacher (Chetty et al., 2011, p. 48). This assumption may be questioned.

A 2.04 standard deviation increase in performance might be possible if rankings were stable and rankings in the current year predicted performance in the following year. As Tables 1 and 2 indicate, however, teacher rankings bounce up and down from year to year. A teacher who ranks in the lowest quartile this year is more likely to rank in the upper three quartiles the next year than to remain in the bottom quartile. Conversely, a teacher who ranks in the highest quartile this year is more likely to drop into the bottom three quartiles the next year than to remain in the top quartile. Chetty et al. (2011) concluded that 75% of the variance in rankings is attributable to random measurement error, rather than true differences in teacher performance (p. 49). Other researchers have found that one-third to one-half of the differentials in teacher performance are driven by random measurement error, rather than true differences in teacher performance (see McCaffrey et al., 2009). Thus, a teacher who appears to rank 2.04 standard deviations above another teacher is not likely to maintain that differential the following year, and it would not be appropriate to assume that substituting a high-performing teacher for a low-performing teacher would result in the same differential in performance next year. The view that teacher rankings are stable over time and actual gains in student achievement next year would equal the measured differential in performance this year is not supported by the evidence in Tables 1 and 2. For this reason, it is unlikely that substituting a teacher who performs highly this year would translate into the expected 2.04 standard deviation gain in performance next year. If that gain is not achieved, then the estimated $266,643 gain in lifetime earnings would not be achieved.

After accounting for the lack of stability in value-added estimates, Chetty et al. (2011) found that the $266,643 gain in lifetime earnings drops to $135,000 (equal to $4,770 for each of the 28.3 students taught by the teacher who is replaced) based on one year of data (p. 49). If VAM is used to replace the lowest 5% of all teachers, any gains would be limited to 5% of all students. The policy would translate, in the aggregate, to an average gain in lifetime earnings of $239 per person, averaged across all students.

Chetty et al. (2011) found that the lifetime gain for an entire classroom of students equals $190,000 if three years of data are available but, as noted above in section 2.1, it would not be uncommon for data to be missing in a way that would prevent averaging. For large numbers of teachers, it would be impractical to average their rankings across two or more years (Newton et al., 2010).

2.5 ADEQUATE TEACHER SUPPLY?

Chetty et al. (2011) assume that there is an adequate supply of unemployed teachers who are ready and willing to be hired and would perform at a level that is 2.04 standard deviations above the performance of teachers who are fired based on value-added rankings. Chetty et al. do not justify this assumption with empirical data. The assumption may be questioned. A simple example illustrates that the vacant teaching positions created when low-performing teachers are fired must ultimately be filled with novice teachers whose performance is significantly worse than the performance of experienced teachers (Gordon et al., 2006; Grissmer, Flanagan, Kawata, & Williamson, 2000; Hanushek, Kain, O’Brien, & Rivkin, 2005; Wenglinsky, 2001). The reason that novice teachers must be hired is because there is a teacher shortage (U.S. Department of Education, 2011). In the aggregate, there are more positions than qualified teachers and overall teacher demand is projected to exceed supply by 35% over the next two decades (Gordon et al., 2006).

To simplify, suppose that there are 10 teaching positions in the entire nation. Suppose that nine of the positions are currently filled with teachers (i.e., there is one vacancy). Suppose, further, that value-added methods could be used to reliably identify the lowest-performing teacher (Teacher 9), who performs at a level that happens to be 2.04 standard deviations below the performance of Teacher 1. If Teacher 9 is fired (and not rehired by any other school), a second vacancy is created. Teacher 9 potentially could be replaced with Teacher 1, but this action simply shifts the second vacancy to Teacher 1’s school. The process of teacher substitution may continue but, at the end of the process, Teachers 1–8 remain employed. There are now two vacant teaching positions that can only be filled with novice teachers. This is true whether the novice teachers arrive as fresh graduates from teaching colleges or as individuals previously employed in nonteaching occupations who choose to switch into the teaching profession through the alternative certification path. If one of those vacant positions is filled with a novice teacher, then any gain in the average level of student achievement (across Teachers 1–9) depends entirely on the difference in performance between (the fired) Teacher 9 and the newly-hired novice teacher. If the newly-hired novice teacher outperforms Teacher 9, then there is a gain in performance; if the novice performs worse than Teacher 9, there is a loss.

One might ask why low-performing teachers cannot be replaced with experienced teachers who leave the teaching force temporarily, then rejoin at a later date. Suppose, for example, that 2 experienced teachers rejoin the teaching force every year. Why is it not possible for those teachers to fill the two vacant teaching slots?

This is only possible if there is no teacher shortage. If the supply of teachers to the profession equals the number of vacancies, then no teacher shortage exists. A shortage can only exist if the supply of teachers is less than the number of vacancies. In the example given above, if it is the case that 2 experienced teachers rejoin the teaching force every year, then it must be the case that 2 experienced teachers leave the force every year, leaving a single vacancy in the absence of Chetty et al.'s (2011) proposal. If 2 experienced teachers rejoin the teaching force but only 1 teacher leaves, then the single remaining vacancy would be filled by Teacher 10: all 10 teaching positions would now be filled and, therefore, there would be no teacher shortage. Recall, however, that there is currently a teacher shortage, which means that it cannot be true that the net inflow of experienced teachers rejoining the teaching force equals or exceeds the number of vacancies (U.S. Department of Education, 2011).[4] Furthermore, if there is an inflow of novice teachers, the inflow of novice plus experienced teachers must be less than the number of vacancies, if indeed there is a teacher shortage.

Currently, in the absence of Chetty et al.'s (2011) proposal, some of the existing vacancies across the nation are being filled with novice teachers, some are being filled with experienced teachers who rejoin the teaching force, and at least one vacancy remains (because there is a teacher shortage)—implying that any *extra* vacancies created by Chetty et al.'s proposal *must* be filled with novices. There is no other possible source. *In the presence of a teacher shortage, it cannot be the case that any of the extra vacancies created by Chetty et al.'s proposal will be filled with experienced teachers. Ultimately, after the type of shuffling described above, all of the extra vacancies must necessarily be filled with novices. Therefore, any policy that involves firing low-performing teachers must acknowledge that the vacant positions will ultimately be filled with novices, not experienced teachers.*

Significantly, when value-added methods are used to identify low-performing teachers, replacing these teachers with novice teachers can have unexpectedly negative effects. For example, McCaffrey et al. (2009) controlled for student fixed effects and found that a policy of replacing the bottom 40% of all teachers would raise student achievement by 0.04 standard deviations if fired teachers were replaced with teachers performing in the top 60%. However, a more realistic assumption is that replacements are novices whose performance is lower than experienced teachers (Gordon et al., 2006; Grissmer et al., 2000; Hanushek et al., 2005; Wenglinsky, 2001). Under the assumption that fired teachers are replaced with novice teachers, the overall impact on student achievement across all students would be negative 0.055 standard deviations (Yeh, 2012). The poor result is a direct consequence of the lack of stability in teacher rankings. The use of value-added methods is unreliable in identifying the bottom 40% of all teachers; when those methods are employed, many teachers who do not "belong" in the low-performing category are fired, while many teachers who do not "belong" in the high-performing category are retained. The result is a very small gain in aggregate performance that is completely offset by the well-established decrease in performance when large numbers of novice teachers are hired to replace experienced teachers (Gordon et al., 2006; Grissmer et al., 2000; Hanushek et al., 2005; Wenglinsky, 2001).

Chetty et al.'s (2011) main analysis excluded the impact of replacing fired teachers with novices. However, in a footnote, they estimated that the students of novice teachers score 0.03 standard deviations below the students of experienced teachers (p. 48). This would reduce the previously estimated 0.0843 standard deviation increase in student test scores to 0.0543 standard deviations for every one standard deviation increase in teacher quality (p. 24). The reduction in impact is significant but smaller than alternative estimates. For example, Gordon et al. (2006) estimated that the average "value-added" of novices is about four percentile points lower than teachers with two years of experience, equal to a negative effect size of 0.171 standard deviations (see the authors' footnote 7 for conversion of percentile point scores into standard deviation units). This would reduce Chetty et al.'s estimated 0.0843 standard deviation increase in student test scores to negative 0.0867 standard deviations. Thus, average student achievement would decrease by 0.0867 standard deviations as a consequence of replacing low-performing teachers with novice teachers. Additional research is needed to determine if the measured impact of novice teachers is artificially depressed by unstable teacher rankings.

The negative effect of replacing low-performing teachers with novice teachers would decrease as novice teachers gain experience, but any argument that long-term gains would be positive is contingent on the reliability and stability of teacher rankings. There is no empirical evidence that long-term gains are positive, and there is no evidence that long-term gains would outweigh the immediate losses that are incurred when novice teachers replace experienced teachers.

2.6 PERSISTENT EFFECTS?

Chetty et al. (2011) found that a teacher's impact fades over time, but one-third of the impact persists; however, other researchers employing stronger analytical methods found that the fade-out is large and quick, and any persistent effect is small. For example, Kane and Staiger (2008) employed random-assignment of teachers to students and found that half of a teacher's impact fades after one year, and an additional 50% fades after the second year, implying that no more than 25% of a teacher's impact persists after two years (see also Carrell & West, 2010; Goldhaber & Hansen, 2012; Jacob, Lefgren, & Sims, 2010; Konstantopoulos, 2011; McCaffrey et al., 2009; Rothstein, 2010).

Therefore, Chetty et al.'s (2011) ad hoc assumption that the 0.9% increase in income observed at age 28 would persist at every age throughout an adult's life, resulting in a cumulative lifetime gain of $4,600, may be questioned (pp. 39, 48). The assumption is not consistent with the evidence that a teacher's impact quickly fades. Perhaps a more reasonable assumption, one that is more consistent with the evidence regarding fade-out, is that the 0.9% increase in income observed at age 28 fades by 50% in each subsequent year.

2.7 IS VAM COST-EFFECTIVE?

Chetty et al. (2011) implicitly assume that the use of VAM to identify and replace low-performing teachers is a cost-effective approach for improving student outcomes, where cost-effectiveness is defined by the resulting gain in student achievement for each dollar invested by society. However, two cost-effectiveness studies indicate that VAM is not cost-effective relative to alternative approaches for raising student achievement (Yeh, 2012; Yeh & Ritter, 2009). Both studies suggest that there are large costs to society of implementing any scheme to replace low-performing teachers: the costs to society of educating new teacher college graduates (including their foregone wages), costs incurred by hiring school districts and schools, costs incurred by new teachers, costs incurred by terminated teachers, the reduced output of terminated teachers while learning a new occupation, the opportunity cost of the labor of newly-hired teachers, the costs of adjudicating terminations based on VAM, the cost to raise salaries for all teachers by an amount that would be necessary to attract more individuals to the teaching profession, and the additional cost to implement VAM assessments.[5] These costs would be offset by the output of terminated teachers in new occupations after a period of retraining and job search but would be substantial.

The termination of a single teacher would create net social costs equal to $314,825.57 (Table 3). If the bottom 10% of all teachers were terminated each year, the annual cost averaged over all teachers would equal (.1) X $314,825.57 or $31,482.56 per teacher. The annual cost per student equals $1,574.13, assuming 20 students per teacher.

The largest cost to society is the opportunity cost of replacing terminated teachers with newly-minted college graduates who obtain teaching certification after one additional year of college coursework. The cost to society includes the value of their foregone output in the next best use of their labor. This may be imputed based on the average beginning teacher salary of $40,049 (U.S. Department of Education, 2005). The present value of this stream over the expected career duration of a new teacher (9.11 years), adjusted for a total compensation-to-salary ratio of 1.43 and assumed to grow at 2% per year (including increases in real income as living standards rise over time as well as seniority-related increases in compensation) but discounted at 5% per year for the present value calculation, is $456,082.06.[6]

This cost to society is offset by the gain in the output of the terminated teachers once they have been retrained and have transitioned into new occupations. While it is not possible to know exactly what occupations the former teachers will transition into, it is reasonable to assume that they will be occupations that require the same level of education (a college degree) and provide roughly the same value of output as teaching. Assuming that retrained workers start in a new occupation at a salary equivalent to a new teacher's salary of $40,049, assuming a compensation-to-salary ratio of 1.43, assuming that wages grow at 2% per year (including increases in real income as living standards rise over time as well as seniority-related increases in compensation), but discounted at 5% per year for the present value calculation, the gain in output to society equals $414,934.59. The income stream begins after an average of 27.36 weeks of retraining (Congressional Budget Office, 1994) and an average of 10.4 weeks to find a new position (Gottschalck, 2006), lasts a period of 8.38 years, and ends 9.11 years after the date of termination. Thus, the income stream is calculated over the same overall time period as the average duration of the new teacher's expected teaching career.

Society would also incur the costs of adjudicating any disputed terminations. Unlike the proposal by Gordon et al. (2006) to use VAM to identify and fire the bottom quartile of novice, untenured teachers (approximately 2% of all teachers), the proposal that is the focus of the current analysis would involve firing a larger percentage of all teachers, a majority of whom would necessarily be tenured teachers who could not be fired without adequate cause. As previously noted, VAM is less reliable than flipping a coin for the purpose of categorizing high- and low-performing teachers (see endnote 2). Thus, the use of VAM to terminate teachers is likely to result in an avalanche of lawsuits by terminated teachers. The evidence overwhelmingly favors litigants who assert that results based on VAM do not meet the legal standard of adequate cause for termination, suggesting that terminated teachers would be likely to win almost every case, since it would be nearly impossible for school districts to show "adequate cause" for termination based on VAM. Districts would have to fall back on existing methods for identifying poor teachers, which currently result in the involuntary termination of a very small percentage of all teachers. In New York, for example, only 88 out of approximately 80,000 city schoolteachers lost their jobs for poor performance over a three-year period—a rate of 0.037% per year (NY Daily News, 2010). In Los Angeles, only 112 of 43,000 tenured teachers faced termination between 1995 and 2005, a rate of 0.026% per year (MSNBC, 2008). In New Jersey, 47 of 100,000 teachers were fired over a 10-year period, a rate of 0.005% per year (MSNBC, 2008). The annual termination rate is 0.01% in Chicago, 0.04% in Cincinnati, and 0.01% in Toledo (Weisberg, Sexton, Mulhern, & Keeling, 2009). In Akron, OH; Denver, CO; Elgin, IL; Jonesboro, AR; and Pueblo, CO; no teachers were formally dismissed over periods that ranged from two to four years (Weisberg et al., 2009). Even if all of these terminated teachers are drawn from the bottom 10% of all teachers subject to termination based on VAM, only small percentages of the VAM-based terminations could be justified based on methods that are independent of value-added rankings: 1.1% in New York, 2.6% in Los Angeles, 0.47% in New Jersey, 0.1% in Chicago, 0.4% in Cincinnati, and 0.1% in Toledo. This implies that litigants who were terminated on the basis of VAM might be expected to prevail in over 97.4% of all cases, assuming that judges agree with the National Academies' Board on Testing and Assessment, which concluded that it is not appropriate to use VAM to make operational decisions regarding teacher hiring and firing (Haertel, 2009, p. 10). As a consequence, and as a consequence of provisions that require teachers to receive their normal pay during the termination process, school districts could expect nearly every termination to be challenged, resulting in enormous costs (Blacher, 2006).

The cost of litigation is high, regardless of the outcome. Tenured teachers often must be provided with names of witnesses, the power of subpoena to compel production of documents and testimony of witnesses, the right to counsel at all stages of the process, and the right to appeal (Blacher, 2006; Nixon, Douvanis, & Packard, 2009). It is estimated that the average cost of terminating a teacher in California is approximately $200,000 (Blacher, 2006).[7] In San Diego, a single termination proceeding took more than four years and cost more than $300,000 in legal fees (Blacher, 2006). In New York, section 3020-a of the state Education Law allows a tenured school district employee who has been charged with incompetence or misconduct to request that a hearing officer review the district's charges and make findings of fact and recommendations as to penalty or punishment, if warranted. In general, "the cost and time required to terminate a permanent teacher are extreme" (Blacher, 2006).

On average, a full 3020-a hearing costs New York districts $216,588 and takes 502 days, according to a New York State School Boards Association survey of 400 districts from 2004 to 2008 (Gould, 2009). This survey provided a breakdown of costs that permits adjustments to reflect the true social costs. The largest expense was the salary and fringe benefits paid to the suspended employees, accounting for 52% of costs. Salaries and benefits for substitute teachers represented 30% of the costs, while legal expenses represented 12% of the costs. Other expenses included other staff costs (5%) and miscellaneous costs, such as the cost of outside investigators, expert witnesses, transcription, photocopying, and travel (1%). However, since the salary and benefits of the suspended employees would have been paid in the absence of the disciplinary hearings, I reduced the total cost figure by 52% to reflect the real social cost incurred by each district, equal to $103,962.24.

In addition to the costs incurred by each district, the suspended employees (or their unions) incurred legal expenses that may be expected to average approximately half the legal expenses incurred by their school districts, equal to $12,995.28 per case. The total social cost of each hearing equals $103,962.24 plus $12,995.28, or a total of $116,957.52 per terminated teacher. This excludes psychic costs incurred by terminated teachers, as well as the cost of any appeals, which could double the cost.

The annual cost of implementing a value-added assessment system may be estimated from the costs of administering and scoring the assessments for Tennessee's Value-Added Assessment System (TVAAS): $5.60 per student, adjusted for inflation and including the cost of the TVAAS reports (Bratton, Horn, & Wright, n.d.).

In addition to the cost of the assessments, salaries must be raised for all teachers in order to attract more individuals to the profession of teaching. This cost is above and beyond the cost to educate and train the new teachers, since there is no army of unemployed teachers waiting to fill the empty teaching slots. On the contrary, there are shortages in many subject specialties and overall teacher demand is projected to exceed supply by 35% over the next two decades (Gordon et al., 2006; U.S. Department of Education, 2011).

The increase in teacher salaries required to attract a sufficient number of new individuals to the teaching profession may be estimated using conservative assumptions. Suppose, for example, that VAM is used to identify and replace the bottom 10% of all teachers. If Q equals the annual supply of teachers, the value-added proposal implies that Q must be increased by 11.11% to an amount equal to 1.11Q (elimination of the bottom 10% of teachers reduces 1.11Q to Q). The cost is determined by the elasticity of teacher supply, defined as the percentage change in the annual supply of new teachers for every one percent change in average annual teacher salary: $\%\Delta Q / \%\Delta S$. I assumed a supply elasticity of three, which is near the top of the range of ordinary supply elasticities estimated by Manski (1987). However, the correct elasticity is likely to be lower, because the use of value-added methods to fire the bottom 10% of all teachers increases the risk of being fired, making teaching a less desirable career choice. Thus, a 1% salary increase is likely to be insufficient to induce a 3% increase in the supply of new teachers, implying that the estimate of the required increase in teacher salaries is likely to be a lower-bound estimate of the true cost.

A supply elasticity of three implies that teacher salaries must increase by 3.70% (11.11/3) to elicit the number of new teachers required to replace the bottom 10% of all teachers. Assuming an average teacher salary of $51,055.19 per year after adjusting for inflation (National Center for Education Statistics, 2005), a compensation-to-salary ratio of 1.43 (U.S. Department of Labor, 2008), and assuming 20 students per teacher, it would cost an extra $135.07 per student per year to raise salaries sufficiently to attract the teachers necessary to replace the bottom 10% of all teachers.

The total annual cost of implementing this proposal is the cost to society of replacing a terminated teacher through a fifth year teacher education program ($1,574.13), plus the cost of the assessments ($5.60), plus the cost to raise salaries sufficiently to replace the bottom 10% of all teachers ($135.07), or a total of $1,714.80 per student. This figure is underestimated to the extent that fired teachers incur psychic losses and to the extent that the increased occupational risk of entering the teaching profession that is implied by firing 10% of all teachers each year would drive teacher salaries upward, raising the cost of hiring new teachers as well as the cost of employing existing teachers.

2.7.1. COST-EFFECTIVENESS RESULTS

If terminated teachers are replaced with average teachers, the average gain across all students is 0.00843 standard deviations per year, and the effectiveness-cost ratio, equal to 0.00843 divided by $1,714.80, is very low: 0.000005. As noted above, however, terminated teachers must be replaced with novice teachers, suggesting that the 0.000005 effectiveness-cost ratio is overestimated. If terminated teachers are replaced with novice teachers whose students perform 0.03 standard deviations below the students of experienced teachers, the average gain across all students is 0.00543 standard deviations per year, and the effectiveness-cost ratio falls to 0.000003. If terminated teachers are replaced with novice teachers whose students perform 0.171 standard deviations below the students of teachers with two years of experience, the average gain across all students is negative 0.00867 standard deviations per year, and the effectiveness-cost ratio is negative 0.000005, implying that student achievement falls by 0.000005 standard deviations for every dollar that is spent to replace low-performing teachers with novice teachers. This result is consistent with previous research suggesting that the overall impact of replacing low-performing teachers is negative (Yeh, 2012).

This result does not change even if the analysis is limited to the subset of students who would benefit from the intervention. If terminated teachers are replaced with novice teachers whose students perform 0.03 standard deviations below the students of experienced teachers, the average gain for these students is 0.0543 standard deviations per year, the cost per student is $17,148, and the effectiveness-cost ratio equals 0.000003. If terminated teachers are replaced with novice teachers whose students perform 0.171 standard deviations below the students of teachers with two years of experience, the average gain for these students is negative 0.0867 standard deviations per year, and the effectiveness-cost ratio is negative 0.000005.

Nor does the result change if the analysis is limited to the 5% subset of students who would benefit under Chetty et al.'s (2011) proposal to use VAM to replace the bottom 5% of all teachers. If terminated teachers are replaced with novice teachers whose students perform 0.03 standard deviations below the students of experienced teachers, the average gain for these students is 0.0543 standard deviations per year. The cost per student in this subset is slightly less ($17,133.85) because of the reduced need to raise teacher salaries to attract a smaller number of teacher replacements to the profession. However, the cost of raising those salaries is concentrated in this subset of students. The effectiveness-cost ratio remains essentially unchanged because this ratio is calculated on a per-student basis and the bulk of the costs are already limited to the teachers and students who would be affected by the proposed policy (see Table 3). Dividing the effect size by the cost per student in this subset produces an effectiveness-cost ratio equal to 0.000003. If terminated teachers are replaced with novice teachers whose students perform 0.171 standard deviations below the students of teachers with two years of experience, the average gain for these students is negative 0.0867 standard deviations per year, and the effectiveness-cost ratio is negative 0.000005.

To determine whether teacher replacement is a cost-effective strategy, it is necessary to compare the approach with other strategies. With regard to the field of education, a cost-effective intervention may be defined as the approach that offers the largest impact with regard to student achievement in math and reading for each dollar invested by society in that intervention (Levin, 1988). Using this definition, teacher replacement is not cost-effective. The effectiveness-cost ratio for rapid performance assessment, an alternative strategy for improving student outcomes, ranges from 0.017152 to 0.028571 (Yeh, 2010a) and is approximately 5,700 times larger than the ratio for Chetty et al.'s (2011) VAM-based teacher replacement strategy (0.000003), implying that Chetty et al.'s strategy is not a cost-effective approach for raising student achievement.

2.7.2. BENEFIT-COST RESULTS

With regard to earnings, Chetty et al.'s (2011) proposed intervention does not meet the test of a benefit-cost analysis. As indicated in section 2.3, if VAM is used to replace the lowest 10% of all teachers, a one standard deviation increase in teacher quality in a single grade is associated with an average gain in lifetime earnings of $460 per person, averaged across all students (in 2010 dollars, Chetty et al., 2011, p. 20). Since the cost calculations for Chetty et al.'s proposed intervention assumed 2006 as the base year, and to ensure comparability with previous cost calculations for performance feedback and other interventions (Yeh, 2010a), I adjusted the $460 figure (using the Consumer Price Index) to 2006 dollars (Bureau of Labor Statistics, 2012). The resulting figure ($425.29), divided by the cost per student (averaged over all students) to implement this intervention for one year ($1,714.80), produces a benefit-cost ratio equal to 0.25. Society would gain $0.25 for every dollar invested in the intervention, implying that the costs of the intervention exceed the benefits by a ratio of four to one.

Chetty et al.'s (2011) proposed intervention does not meet the test of a benefit-cost analysis even when limited to the 10% subset of students who would benefit from the intervention. In this subset, a one standard deviation increase in teacher quality in a single grade is associated with an average gain in lifetime earnings of $4,600 per student. Adjusted to the 2006 base year, the resulting figure ($4,252.85), divided by the cost per student in this subset ($17,148), equals a benefit-cost ratio of 0.25, implying that the costs of the intervention remain four times greater than the benefits.

Chetty et al.'s (2011) proposed intervention does not meet the test of a benefit-cost analysis even when limited to the subset of students who would benefit if VAM is used to replace the lowest 5% of all teachers. As indicated in section 2.3, a 2.04 standard deviation increase in teacher quality in a single grade is associated with an average gain in lifetime earnings of $9,422 per person. Adjusted to the 2006 base year, the resulting figure ($8710.95), divided by the cost per student in this subset ($17,133.85), equals a benefit-cost ratio of 0.51, implying that the costs of the intervention are almost twice the benefits.[8] If tenure is eliminated and all teachers are employed at-will, litigation costs may be excluded; the cost per student falls to $11,285.98, but the benefits of the intervention ($8710.95) remain smaller than the costs. This analysis also applies to the case where the proposed policy is only applied to novice teachers who are not tenured. If the policy is applied to a mixture of experienced and novice teachers, the policy would have an effect that falls between the two estimates but, as indicated, the policy does not meet a benefit-cost test under either scenario. This negative result holds whether the proposed policy is implemented once or on an ongoing basis, because the costs are incurred every time the policy is implemented and presumably the benefits are received every time, so the ratio of benefits to costs (or the ratio of effect size to costs) would remain unchanged.

3. DISCUSSION

The literature reviewed in Section 1 together with the preceding analysis suggests that the use of value-added statistical methods to identify and replace low-performing teachers is not warranted. VAM lacks sufficient reliability and validity for the purpose of hiring and firing teachers. Once gains are averaged over all students, they would be very small. Furthermore, it appears that any gains would fade away very quickly. Significantly, the approach is neither cost-effective nor does it meet the test of a benefit-cost analysis.

While the preceding analysis is based on Chetty et al. (2011), much of the analysis applies to any proposal to use value-added methods to replace low-performing teachers. Studies of the stability of VAM-based teacher rankings have found inadequate reliability for operational decisions regarding the hiring and firing of teachers (Haertel, 2009). Even when studies of VAM are taken at face value, the results indicate small impacts on student achievement (Chetty et al., 2011; Goldhaber & Hansen, 2012; Gordon et al., 2006; McCaffrey et al., 2009). When these results are integrated with analyses of the full social costs of implementing VAM in order to replace low-performing teachers, it becomes clear that VAM is not cost-effective relative to the most promising strategies for raising student achievement (Yeh, 2010a, 2012; Yeh & Ritter, 2009).

These results suggest a need to revisit the assumption that large improvements in student outcomes may be achieved by identifying and replacing low-performing teachers. This assumption suggests that high- and low-performing teachers are analogous to "good apples" and "bad apples" and implies that average teacher quality would improve if we got rid of bad apples. The assumption is that teacher quality is a fixed characteristic—that a high-performing teacher this year will be a high-performing teacher next year, and a low-performing teacher this year will be a low-performing teacher next year. As indicated in section 2.1, this assumption is not supported by the available data. Teacher quality is not a fixed, inherent characteristic but instead fluctuates over time and is variable in a way that is not captured by a model that categorizes workers as "good apples" and "bad apples" (Goldhaber & Hansen, 2012). Much of a teacher's performance varies over time due to unobservable factors such as effort, motivation, and class chemistry that are not easily captured through VAM (Goldhaber & Hansen, 2012).

Advocates of using VAM for high-stakes decisions regarding teacher hiring and firing argue that an excessive concern with false identifications of low-performing teachers serves the interests of teachers rather than their students (Glazerman, Loeb, et al., 2010). Framing the issue in this way, however, sets up a false dichotomy. The question is not whether society should serve the interests of teachers rather than their students, or what is the proper balance between false positive and false negative identifications, but *what is the most efficient approach for raising student achievement?* A number of cost-effectiveness analyses have now been performed that permit comparison of 22 of the leading approaches for raising student achievement (Yeh, 2010a). The results from section 2.7 suggest that the most efficient approach—rapid performance feedback—is approximately 5,700 times as efficient as the use of VAM to identify and replace low-performing teachers.

This result may appear to be improbable. There are two reasons for the tremendous disparity in efficiency. First, the particular variant of rapid performance feedback that is the focus of the comparison (the Accelerated Reader and Accelerated Math programs, collectively labeled "rapid assessment" programs) involves changes in the way learning material is individualized and presented to students in combination with performance feedback in the form of individualized daily assessments. These changes apparently alter students' perceptions of their abilities to improve their performances, so that low-performing students begin to believe that they can achieve academic success through their own efforts (Yeh, 2006). Students appear to acquire an internal locus of control, exerting more effort than students who do not receive rapid performance feedback (Yeh, 2006, 2010b). This approach offers a different way of thinking about how student performance may be improved. In contrast, VAM-based teacher replacement policies attempt to improve student achievement without addressing the psychology of student learning.

A second reason for the disparity in efficiency between VAM and rapid assessment is that rapid assessment is primarily implemented with the aid of computer software, the cost of which can be amortized over multiple years and spread over hundreds of students in each school building. The annual cost per student is very low. In contrast, as indicated by the analysis in section 2.7, the use of VAM to identify and replace low-performing teachers is tremendously costly.

In contrast to the rapid performance feedback model, the use of VAM to identify and replace low-performing teachers relies on the conventional model of instruction, which fails to individualize task difficulty and therefore fails to change the tedious experience of schooling for students who are above-average and the discouraging experience of schooling for students who are below-average. Failing to address these dynamics, VAM-based policies place the entire burden of raising student achievement on teachers who are locked into systems that appear to inadvertently undermine student engagement and achievement. As indicated above, VAM-based teacher replacement policies are approximately 5,700 times less cost-effective than Accelerated Reader or Accelerated Math, suggesting that VAM-based teacher replacement is not a cost-effective approach for raising student achievement.

Notes

1. The version of the Chetty, Friedman, and Rockoff (2011) study that is analyzed here is posted at http://obs.rc.fas.harvard.edu/chetty/value_added.pdf. The study was conducted under the auspices of the National Bureau of Economic Research (NBER), the nation's leading nonprofit economic research organization. NBER publishes rigorous economic analyses from leading scholars prior to their publication in academic journals.

2. Suppose that half of a sample of teachers is fired, using a coin flip to determine the fate of each teacher. From Table 1, a minimum of 60% of all teachers deserve to be retained, while 40% do not, according to the year $t+1$

teacher rankings, even when the sample is drawn from the bottom quartile, as determined by year *t* VAM rankings. The coin flip results in firing half of those who deserve retention (30% of all teachers) and retention of half of those who deserve firing (20% of all teachers) for an overall error rate of 50%. In comparison, when VAM is used to identify and fire the bottom quartile (or bottom quintile) of teachers, the results of Tables 1 and 2 imply that this decision is incorrect, according to the year $t+1$ teacher rankings, for a minimum of 59% of the teachers in that quartile (or quintile), for an overall error rate of 59%. Thus, a VAM-based decision rule is less reliable than flipping a coin.

3. The validity of using test scores for a particular purpose depends on "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985, p. 9). Validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). In the case of teacher rankings based on value-added test scores, the inference that the results reliably categorize teachers as either high-performing or low-performing teachers is not appropriate, nor does the available evidence support the use of value-added teacher rankings for the purpose of high-stakes decisions regarding hiring, firing, promotion, or compensation.

4. An oversupply of teachers in large urban districts that are reducing their teaching forces may permit experienced teachers to be hired to replace low-performing teachers who are terminated.

5. It may be argued that low-performing teachers are not well-matched to the occupation of teaching and, therefore, there would be a gain to society if those individuals are redirected to other occupations. However, the numerator of the benefit-cost ratio (see section 2.7.2) accounts for this gain to society, measured in terms of the increase in the lifetime earnings of students taught by teacher replacements who are presumably better suited to the occupation of teaching than the terminated teachers. The denominator of the benefit-cost ratio accounts for the costs of teacher replacement. In addition, the hypothesis that fired teachers are better suited to other occupations is not supported by the available evidence. Only 3.8% of new female elementary teachers and 5.4% of new female high school teachers who left full-time teaching during the 1994-2001 time period took a non-education-sector job in Georgia that paid more than the state minimum teaching salary in Georgia (Scafidi, Sjoquist, & Stinebrickner, 2006). Since these figures include all exiting teachers, including teachers who left voluntarily and, therefore, were likely to be considered more productive by potential employers than teachers who were fired, it is likely that the percentage of fired teachers who took non-education-sector jobs paying more than the state minimum teaching salary is even lower. This implies that well over 94% of fired teachers are unable to earn more in their new occupations. Fired teachers are not more productive in new occupations.

6. Note that the 3% discount rate used by Chetty et al. (2011) to discount earnings (see section 2.3, above) is based on their assumption of 2% wage growth adjusted for a 5% discount rate (p. 39). To ensure consistency, I use the same assumptions for wage growth and the discount rate as Chetty, et al. These assumptions are slightly different than the assumptions used in Yeh and Ritter (2009) and Yeh (2012).

The best available estimate of the career duration of the average teacher was derived using proportional hazards modeling, which accounts for the difficulty of estimating career duration when some members of the research sample have not exited the teaching profession by the end of the research study period (Murnane, Singer, & Willett, 1988). Proportional hazards modeling incorporates information about the pattern of teacher attrition during the study period to predict the median length of each spell of teaching. Using data from Michigan covering a 12-year time period, Murnane et al. provided separate estimates for six subject area specialties of the duration of the average teacher's first two spells of teaching. The authors reported the percentage distribution of teachers across the six subject area specialties as well as the percentage of teachers in each of the six subject areas who returned to teaching after a career interruption. I used this information to calculate the average career duration (9.11 years) for an average teacher, weighted by the percentage distribution of teachers across the six subject area specialties and including the expected length of a second spell of teaching based on the probability of a second spell.

7. Terminating a teacher using VAM would likely be even more litigious and costly if a judge agrees with the judgment of the National Research Council's (NRC) Board on Testing and Assessment, which concluded that VAM is not sufficiently reliable for the purpose of terminating teachers (Haertel, 2009). The NRC's judgment that VAM is unreliable is independent of the author's view that VAM is unreliable.

8. These calculations assume that Chetty et al.'s (2011) proposal is implemented on an ongoing annual basis. The benefits and costs are calculated per student in each cohort that would benefit from the intervention. However, it might be argued that each replacement teacher generates a "legacy" stream of student cohorts that benefit from the increased productivity of that teacher, and, therefore, the benefit of replacing each teacher should be multiplied by the number of cohorts taught by each teacher. This might be accurate if no fired teacher was rehired, if the composition of the teaching force was frozen and there were no retirements or exits by any teacher in the entire teaching force, and if annual culling of the teaching force reliably eliminated low-performing teachers. To clarify:

a. If all fired teachers were rehired by other schools, the benefit of Chetty et al.'s proposal would drop to zero, but most of the costs would remain.

b. If some fired teachers were rehired by other schools, the benefit of Chetty et al.'s proposal would diminish in proportion to the degree of rehiring.

c. In principle, the ability of the federal government to regulate policies regarding the rehiring of fired teachers is limited because the U.S. Constitution effectively delegates this role to the states.

d. While the federal government might be able to require, as a condition of receiving federal education funds, that each state must implement a policy forbidding the rehiring of a fired teacher, it is likely that such a regulation would be widely opposed for two reasons: 1) If VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 1 and 2 indicate that this decision is incorrect, according to the year $t+1$ teacher rankings, between 59 and 70% of the time. 2) There is a teacher shortage and many classrooms could not be staffed under such a requirement.

e. In the absence of a federal requirement forbidding the rehiring of fired teachers, each state would establish its own policy. The benefit of Chetty et al.'s proposal would depend on each state and/or locality establishing such a requirement. It is likely that such a regulation would be widely opposed for both of the reasons specified in d., above.

f. However, even under the strong assumption that no state permits the rehiring of fired teachers, the dynamics of the teacher labor market would cause the legacy benefits of Chetty et al.'s proposal to fade every year, unless the culling process envisioned in Chetty et al.'s proposal is implemented on an ongoing annual basis. The reason is that a significant portion of the entire teaching force exits every year and is replaced by a new set of teachers with heterogeneous abilities. This "waters down" the composition of the upper 95% that was retained under Chetty et al.'s policy with novices representing the full distribution—the full bell curve—of teacher ability. To understand the issue, consider an extreme example: Assume that attrition is 100% after one year, and all teachers are replaced with novices. Clearly, there would be no legacy benefit of Chetty et al.'s proposal, because the distribution of teacher performance would reflect the entire bell curve from that point forward. If attrition is 50%, then the benefits of the proposal are cut in half. Benefits are reduced even if attrition is limited entirely to the upper 95% of all teachers because attrition and replacement cause that group of teachers to take on the characteristics of the full distribution of teachers—the entire bell curve, not just the upper 95% of the bell curve. The only way to maintain the upper 95% advantage that is presumably conferred by Chetty et al.'s proposal is through ongoing culling. Data from the national Schools and Staffing Survey indicate teacher retention rates of 76% after two years, 67% after three years, 60% after four years, and 54% after five years (Quartz et al., 2004). These figures include retention in all roles within the field of education, not only teaching, implying that the teacher retention rate is lower and attrition is a significant problem.

g. Would annual culling gradually improve the stock of teachers over time, overcoming slippage due to attrition? This depends on the validity of the assumptions underlying Chetty et al.'s analysis. In particular, if VAM is used to identify and fire the bottom quartile (or quintile) of teachers, the results in Tables 1 and 2 indicate that this decision is incorrect, according to the year $t+1$ teacher rankings, for 59 to 70% of the teachers. These results suggest that productive teachers would be culled more frequently than unproductive bottom quartile (or bottom quintile) teachers. The problem is illustrated by data from six large urban school districts indicating that an English language arts teacher who is predicted, based on VAM, to score at the 25th percentile is actually more likely to fall in the top half of the distribution than in the bottom quarter (Rothstein, 2011, pp. 8-9). Depending on the distribution of teachers, this implies the possibility that a VAM-based decision rule to fire the bottom quartile of teachers could actually reduce the quality of the teaching force! A second implication is that it would be extremely difficult to justify any policy prohibiting the rehiring of fired teachers, if, in fact, teachers who are predicted to score at the 25th percentile are actually more likely to score above average. Third, it raises serious doubts about not only the validity of Chetty et al.'s analysis but also all proposals to use VAM-based decision rules to fire low-performing teachers. The evidence affirms the conclusion of the NRC's expert panel that VAM is not sufficiently reliable to make operational decisions about firing teachers (Haertel, 2009).

h. Given the evidence in sections 2.1, 2.2, 2.4, 2.5, and 2.6 of this article, plus the strong likelihood that many (perhaps most) fired teachers would be rehired (because prohibitions against rehiring would be difficulty to justify), plus the slippage of gains due to attrition, and, most importantly, the evidence that VAM-based teacher rankings fluctuate up and down and are poor predictors of future performance, both the short- and long-term benefits of Chetty et al.'s policy are questionable. Using VAM, a teacher who is ranked "poor" this year is more likely to be classified as a productive teacher next year than to remain a poor teacher. Therefore, the gains from replacing "poor" teachers are questionable and it would be questionable to multiply those presumed gains by the number of cohorts taught by each teacher. In any case, the average career of a teacher is 9.11 years (see endnote 5); multiplying the effectiveness-cost ratio for Chetty et al.'s intervention by 9.11 gives a ratio equal to 0.000027 (9.11 X 0.000003), which remains 600 times less cost-effective than the ratio for rapid performance feedback. Even after accounting for the legacy benefits of Chetty et al.'s policy, it is far less cost-effective than performance feedback. While additional research is needed to clarify each of these issues, the burden is on advocates to demonstrate that VAM-based teacher replacement is a cost-effective strategy, compared to rapid performance feedback and other leading alternatives.

References

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed.), *Value added models in education: Theory and applications* (pp. 1-26). Maple Grove, MN: JAM Press.

Ballou, D. (2012). *Review of the long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. Boulder, CO: University of Colorado, National Education Policy Center.

Ballou, D., & Podgursky, M. (1995). Recruiting smarter teachers. *The Journal of Human Resources, 30*(2), 326-338.

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

Blacher, M. (2006). K-12 teacher termination hearings: Are they worth the cost? *CPER Journal, 180*, 13-19. Retrieved from http://www.lcwlegal.com/newspublications/Articles/PastArticles/CPER_1006_Blacher.htm

Bratton, S. E., Jr., Horn, S. P., & Wright, S. P. (n.d.). Using and interpreting Tennessee's Value-Added Assessment System: A primer for teachers and principals. Retrieved December 1, 2006, from http://www.shearonforschools.com/documents/TVAAS.HTML#Let's

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. National Research Council, Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability. Washington, DC: The National Academies Press.

Briggs, D., & Domingue, B. (2011). *A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center.

Bureau of Labor Statistics. (2012). *Consumer Price Index*. Washington, DC: U.S. Department of Labor. Retrieved from ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409-432.

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.

Colorado State Council for Educator Effectiveness. (2011). *Report and recommendations*. Retrieved from http://www.cde.state.co.us/EducatorEffectiveness/downloads/Report%20&%20appendices/SCEE_Final_Report.pdf

Congressional Budget Office. (1994). An analysis of the administration's proposed program for displaced workers. *CBO Papers* (Vol. 2006, pp. 25). Washington, D.C.: author.

Dillon, S. (2010, September 1). Formula to grade teachers' skill gains acceptance, and critics. *New York Times,* pp. A1, A3.

District of Columbia Public Schools. (2012). *Value-added*. Retrieved August 10, 2012, from http://www.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/Value-Added

Glazerman, S., Goldhaber, D., Loeb, S., Staiger, D. O., & Whitehurst, G. J. (2010). *America's teacher corps*. Washington, DC: The Brookings Institution.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D. O., Raudenbush, S. W., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

Goldhaber, D., & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions* (Working Paper 31). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D., & Hansen, M. (2012). *Is it just a bad class? Assessing the long-term stability of estimated teacher performance* (Working Paper 73). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper 2006-01). Washington, D.C.: The Brookings Institution.

Gottschalck, A. O. (2006). *Dynamics of economic well-being: Spells of unemployment 2001–2003*. Washington, DC: U.S. Census Bureau.

Gould, P. (2009). *3020-a process remains slow, costly*. Retrieved October 22, 2010, from the New York State School Boards Association website: http://www.nyssba.org/index.php?src=news&refno=853&category=On%20Board%20Online%20May%2011%202009

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND Corporation.

Haertel, E. H. (2009). *Letter report to the U.S. Department of Education on the Race to the Top Fund*. Washington, DC: National Research Council, The National Academies Press. Retrieved from https://download.nap.edu/catalog.php?record_id=12780

Hanushek, E. A. (2009a). *Teacher deselection*. Retrieved June 9, 2011, from http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/Hanushek%202009%20CNTP%20ch%208.pdf

Hanushek, E. A. (2009b). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: Urban Institute Press.

Hanushek, E. A. (2010). The difference is great teachers. In K. Weber (Ed.), *Waiting for "superman": How we can save America's failing public schools* (pp. 81-100). New York: PublicAffairs.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

Hess, F. M. (2010, October 18). Beyond school choice. *National Review,* 41-42.

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

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*(4), 520-536.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning gains. *Journal of Human Resources, 45*(4), 915-943.

Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh Public Schools*. Cambridge, MA: Mathematica Policy Research.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Cambridge, MA: National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Columbia, MO: University of Missouri.

Koedel, C., & Betts, J. R. (2011). 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.

Konstantopoulos, S. (2011). Teacher effects in early grades: Evidence from a randomized study. *Teachers College Record, 113*(7), 1541-1565.

Kristof, N. D. (2012a, January 11). The value of teachers. *New York Times*. Retrieved from http://www.nytimes.com/2012/01/12/opinion/kristof-the-value-of-teachers.html?scp=3&sq=chetty%20friedman%20rockoff&st=cse

Kristof, N. D. (2012b, January 13). The value of teachers. *The International Herald Tribune,* p. 7.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34*(1), 109-121.

Levin, H. M. (1988). Cost-effectiveness and educational policy. *Educational Evaluation and Policy Analysis, 10*(1), 51-69.

Louisiana Department of Education. (2011). *Measuring teacher impact on student growth in tested grade and subjects (value-added)*. Retrieved May 3, 2011, from http://www.doe.state.la.us/topics/value_added.html

Lowrey, A. (2012, January 6). Big study links good teachers to lasting gain. *New York Times,* p. A1.

Manski, C. F. (1987). Academic ability, earnings, and the decision to become a teacher: Evidence from the National Longitudinal Study of the high school class of 1972. In D. Wise (Ed.), *Public sector payrolls* (pp. 291-312). Chicago, IL: University of Chicago Press.

Martineau, J. A. (2006). Distorting value-added: The use of longitudinal, vertically-scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62.

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

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

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

MSNBC. (2008, June 28). *Superintendent: Bad teachers hard to fire: Some say teacher tenure rules need to be overhauled to address problem*. Retrieved October 21, 2010, from http://www.msnbc.msn.com/id/25430476/

Murnane, R. J., Singer, J. D., & Willett, J. B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational Researcher, 17*(6), 22-30.

New York State Department of Education. (2011). *New York state teacher and principal evaluation: Summary of provisions in draft regulations*. Retrieved from http://usny.nysed.gov/rttt/docs/summary.pdf

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

Nixon, A., Douvanis, G., & Packard, A. (2009, October). *Negative retention: Why probationary teachers are non-renewed.* Paper presented at the Phi Delta Kappa International 2009 Summit on Quality Educator Recruitment and Retention, Indianapolis, IN.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Quartz, K. H., Lyons, K. B., Masyn, K., Olsen, B., Anderson, L., Thomas, A., . . . Horng, E. L. (2004). *Retention report series: A longitudinal study of career urban educators* (pp. 17). Los Angeles: University of California, Los Angeles.

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

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Behavioral and Educational Statistics, 29*(1), 117-120.

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

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

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

Rothstein, J. (2011). Review of *Learning about teaching*. Boulder, CO: National Education Policy Center.

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

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

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

Scafidi, B., Sjoquist, D. L., & Stinebrickner, T. R. (2006). Do teachers really leave for higher paying jobs in alternative occupations? *Advances in Economic Analysis and Policy, 6*(1, Article 8), 1-42.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review, 23*, 411-430.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives, 24*(3), 97-118.

Tennessee Department of Education. (2011). *TVAAS teacher report (for educators)*. Retrieved from http://www.tn.gov/education/assessment/test_results.shtml

The Center for Greater Philadelphia. (2004). *Value-added assessment*. Retrieved February 8, 2008, from http://www.cgp.upenn.edu/ope_value.html#9

The tenure trap: Be careful when hiring teachers—you may be stuck with them for life. (2010, September 28). *NY Dails News*. Retrieved from http://www.nydailynews.com/opinions/2010/09/28/2010-09-28_the_tenure_trap.html

Turque, B. (2012, May 24, 2012). D.C. teacher evaluation formula could change. *The Washington Post*. Retrieved from http://www.washingtonpost.com/local/education/dc-teacher-evaluation-formula-could-change/2012/05/24/gJQACOyRoU_story.html

U.S. Department of Education. (2012). *Teacher incentive fund*. Retrieved August 15, 2012, from http://www2.ed.gov/programs/teacherincentive/index.html

U.S. Department of Education, National Center for Education Statistics. (2005). *Digest of education statistics*. Retrieved from http://nces.ed.gov/programs/digest/2005menu_tables.asp

U.S. Department of Education, Office of Postsecondary Education. (2011). *Teacher shortage areas: Nationwide listing, 1990-91 thru 2011-12*. Washington, DC: U.S. Department of Education.

U.S. Department of Labor. (2008). *Employer costs for employee compensation: December 2007* [Press release]. Retrieved March 26, 2008, from http://www.bls.gov/news.release/pdf/ecec.pdf

Watanabe, T. (2011, March 28). “Value-added” teacher evaluations: L.A. Unified tackles a tough formula. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2011/mar/28/local/la-me-adv-value-add-20110328

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness* (2nd ed.). Brooklyn, NY: The New Teacher Project.

Wenglinsky, H. (2001). *Teacher classroom practices and student performance: How schools can make a difference* (Research Report RR-01-19). Princeton, NJ: Educational Testing Service.

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

Yeh, S. S. (2006). High stakes testing: Can rapid assessment reduce the pressure? *Teachers College Record, 108*(4), 621-661.

Yeh, S. S. (2010a). The cost-effectiveness of 22 approaches for raising student achievement. *Journal of Education Finance, 36*(1), 38-75.

Yeh, S. S. (2010b). Understanding and addressing the achievement gap through individualized instruction and formative assessment. *Assessment in Education, 17*(2), 169-182.

Yeh, S. S. (2012). The reliability, impact and cost-effectiveness of value-added teacher assessment methods. *Journal of Education Finance, 37*(4).

Yeh, S. S., & Ritter, J. (2009). The cost-effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance, 34*(4), 426-451.

STUART S. YEH is Associate Professor and Coordinator of the Evaluation Studies Program at the University of Minnesota. He has conducted numerous cost-effectiveness evaluations and is the author of *The Cost-Effectiveness of 22 Approaches for Raising Student Achievement*.