Journal of Educational and Behavioral Statistics
http://jebs.aera.net

What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?

Peter Z. Schochet and Hanley S. Chiang

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2013 38: 142
originally published online 10 April 2012
DOI: 10.3102/1076998611432174

Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 2, pp. 142–171
DOI: 10.3102/1076998611432174

What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?

Peter Z. Schochet
Hanley S. Chiang
Mathematica Policy Research

This article addresses likely error rates for measuring teacher and school performance in the upper elementary grades using value-added models applied to student test score gain data. Using a realistic performance measurement system scheme based on hypothesis testing, the authors develop error rate formulas based on ordinary least squares and Empirical Bayes estimators. Empirical results suggest that value-added estimates are likely to be noisy using the amount of data that are typically used in practice. Type I and II error rates for comparing a teacher's performance to the average are likely to be about 25% with 3 years of data and 35% with 1 year of data. Corresponding error rates for overall false positive and negative errors are 10% and 20%, respectively. Lower error rates can be achieved if schools are the performance unit. The results suggest that policymakers must carefully consider likely system error rates when using value-added estimates to make high-stakes decisions regarding educators.

Keywords: *value-added models, performance measurement systems, student learning gains, false positive and negative error rates*

Student learning gains, as measured by students' scores on pretests and posttests, are increasingly being used to evaluate educator performance. Known as "value-added" measures of performance, the average gains of students taught by a given teacher, instructional team, or school are often the most important outcomes for performance measurement systems that aim to identify instructional staff for special treatment, such as rewards and sanctions.

Spurred by the expanding role of value-added measures in educational policy decisions, an emerging body of research has consistently found—using available data—that value-added estimates based on a few years of data can be imprecise. In this article, we add to this literature by systematically examining—from a *design* perspective—misclassification rates for commonly used performance measurement systems that rely on hypothesis testing.

Ensuring the precision of performance measures has taken on greater importance with the proposal and implementation of policies that require the use of value-added measures for higher-stakes decisions. The use of value-added measures for policy decisions has been fueled by expanded financial support from the federal government and private foundations. Under the federal Race to the Top grant program, to which $4 billion has been authorized by the *American Recovery and Reinvestment Act of 2009*, a key criterion for selecting state grantees is that they must use performance measures—based heavily on student gains—to inform decisions on the professional development, compensation, promotion, tenure status, and removal of teachers and principals. Indeed, one of the program's first grantees, Tennessee, will proceed to use value-added measures "for making all critical human capital decisions in [the] state's education system" (U.S. Department of Education, 2010). Similar reforms are being financed by hundreds of millions of dollars from the Bill and Melinda Gates Foundation. With these and other initiatives, the use of value-added measures is likely to become even more widespread in the coming years.

Given that individual teachers and schools can be subject to significant consequences on the basis of their value-added estimates, researchers have increasingly paid attention to the precision of these estimates. A number of studies have examined the extent to which differences in single-year performance estimates across educators are due to *persistent* (or long-run) differences in performance—the types of differences intended to be measured—rather than to transitory student-level and classroom-level influences that induce random error, and thus imprecision, in the estimates (see, e.g., Kane & Staiger, 2002a, 2002b).

Existing research has consistently found that teacher-level and school-level averages of student test score gains can be unstable over time. Studies have found only moderate year-to-year correlations—ranging from .2 to .6—in the value-added estimates of individual teachers (Goldhaber & Hansen, 2008; McCaffrey, Sass, Lockwood, & Mihaly, 2009) or small to medium-sized school grade-level teams (Kane & Staiger, 2002b). As a result, there are significant annual changes in teacher rankings based on value-added estimates (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Goldhaber & Hansen, 2008; Koedel & Betts, 2007; McCaffrey et al., 2009).

While previous work has documented instability in value-added estimates post hoc using several years of available data, the specific ways in which performance measurement systems should be designed *ex ante* to account for instability of the estimates have not been examined. This article is the first to systematically examine this precision issue from a design perspective focused on the following question: "What are likely error rates in classifying teachers and schools in the upper elementary grades into performance categories using student test score gain data that are likely to be available in practice?" These error rates are critical for assessing appropriate sample sizes for a performance measurement system that aims to reliably identify low- and high-performing teachers and schools.

We address this precision question both theoretically and empirically. For the theoretical analysis, we employ a commonly used statistical framework for calculating value-added estimates using ordinary least squares (OLS) and Empirical Bayes (EB) methods, and derive associated variance formulas. We then model a realistic performance measurement system that uses hypothesis testing to classify educators into performance categories, and we use the variance formulas to derive equations for calculating system error rates. The formulas depend on several parameters, and we obtain realistic values for these parameters from a synthesis of several published value-added studies and data from recent school-based evaluations. We then calculate system error rates for various assumed sample sizes by applying the theoretical formulas with these empirically based parameter values.

Statistical Framework for the Teacher-Level Analysis

The Basic Statistical Model and Assumptions

Our analysis is based on standard education production functions that are often used in the literature to obtain estimates of school and teacher value-added using longitudinal student test score data linked to teachers and schools (Harris & Sass, 2006; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rothstein, 2010; Todd & Wolpin, 2003). We formulate reduced-form production functions as variants of a four-level hierarchical linear model (HLM; Raudenbush & Bryk, 2002). The HLM corresponds to students in Level 1 (indexed by *i*), classrooms in a given year in Level 2 (indexed by *t*), teachers in Level 3 (indexed by *j*), and schools in Level 4 (indexed by *k*):

$$\text{Level 1 : Students :} \quad g_{itjk} = \xi_{tjk} + \varepsilon_{itjk} \qquad (1a)$$

$$\text{Level 2 : Classrooms :} \quad \xi_{tjk} = \tau_{jk} + \omega_{tjk} \qquad (1b)$$

$$\text{Level 3 : Teachers :} \quad \tau_{jk} = \eta_k + \theta_{jk} \qquad (1c)$$

$$\text{Level 4 : Schools :} \quad \eta_k = \delta + \psi_k \,. \qquad (1d)$$

In this model, g_{itjk} is the gain score (posttest–pretest difference) for student *i* in classroom (year) *t* taught by teacher *j* in school *k*; ξ_{tjk}, τ_{jk}, and η_k are level-specific random intercepts; ε_{itjk}, ω_{tjk}, θ_{jk}, and ψ_k are level-specific *iid* $N(0, \sigma^2_{\varepsilon})$, *iid* $N(0, \sigma^2_{\omega})$, *iid* $N(0, \sigma^2_{\theta})$, and *iid* $N(0, \sigma^2_{\psi})$ random error terms, respectively, where the error terms across equations are distributed independently of one another; and δ is the expected student gain score in the geographic area used for the analysis—which is assumed to be a *school district* (but, e.g., could also be a state, a group of districts, or a school). Note that the level-specific intercepts are conceptualized as random in this framework, although we sometimes treat some intercepts as fixed.

Although this model uses gain scores as the outcome variable, it is closely related to an alternative model, the "quasi-gain" model, in which posttest scores are the outcome variable and pretest scores are a Level 1 covariate. Let β_1 denote the coefficient on the pretest score in the quasi-gain model. If pretest scores are subtracted from both sides of the quasi-gain model, it becomes a gain-score model in which there is a pretest covariate with a coefficient of $(\beta_1 - 1)$. Under the restriction that $\beta_1 = 1$, the quasi-gain model reduces to our benchmark HLM. Because previous work has found that value-added estimates with and without this restriction are very similar (Harris & Sass, 2006; McCaffrey et al., 2004), we maintain this restriction in our benchmark model to allow for simple, transparent estimator, and variance formulas. In practice, there could be precision gains from permitting β_1 to be unrestricted—which, in a gain-score model, would allow prior achievement levels to account for some of the student-level variation in current gains.[1] Thus, in our empirical analysis, we conduct sensitivity tests that relax the assumption of $\beta_1 = 1$ by including pretest scores as a covariate in the HLM.

Our benchmark HLM does not include any other student-level or teacher-level covariates. The bulk of the evidence indicates that demographic characteristics explain very little of the variation in test score gains (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007). Again, we thus omit covariates to permit straightforward formulas of the considered estimators and variances.

Estimates of τ_{jk} in the HLM model are the focus of the teacher-level analysis. A τ_{jk} is interpreted as the expected gain score of a randomly chosen student if assigned to teacher j. We follow the previous literature in assuming that τ_{jk} is constant (persistent) during the evaluation period, so this framework ignores dynamic growth in teacher performance over time. As can be seen by inserting (1d) into (1c), $\tau_{jk} = \delta + \psi_k + \theta_{jk}$, so τ_{jk} reflects (a) the contribution of all educational and background inputs influencing the expected student test score gain in the district (δ); (b) the contribution of factors common to all teachers in the same school (ψ_k), such as the influence of the principal, school resources, and the sorting of true teacher quality across schools; and (c) the contribution of the teacher net of any shared contribution by all teachers in her school (θ_{jk}). As can be seen further from (1a) and (1b), g_{itjk} is influenced not only by τ_{jk} but also by a random transitory classroom effect ω_{jkt} (e.g., a particularly disruptive student in the class), and by a random student-level factor ε_{ijkt}.

In this article, we consider performance schemes that compare estimates of τ_{jk} for all upper elementary school teachers in a hypothetical school district, including those from different schools. We assume that teachers teach self-contained classrooms, where each teacher is assumed to teach a single classroom per year. For notational simplicity, we assume a balanced design, where data are available for c self-contained classes per teacher (i.e., for c years, so that $t = 1, \ldots, c$) with n new students per class each year (so that $i = 1, \ldots, n$) and m teachers in each of the s schools (so that $j = 1, \ldots, m$). For unbalanced designs, the formulas presented in this article apply approximately using mean values for c, n, and m (Kish, 1965).

We focus on the upper elementary grades because there is available empirical evidence on key parameters affecting the precision of value-added estimates, and pretests are likely to be available for analysis. We note, however, that more precise value-added estimates could be obtained for middle school teachers who teach multiple classes per year.

To permit a focused and tractable analysis, we assume a best-case scenario in which commonly cited difficulties with value-added estimation are resolved or nonexistent. In particular, we assume vertical test scales (to allow comparisons across grades), no teacher mobility, and random assignment of students to classrooms and schools (to allow for unbiased estimation of differences in τ_{jk}). Assuming this best-case scenario is likely to produce lower bounds for the error rates that performance measurement systems used in practice are expected to generate.

The HLM used for the analysis can be considered as modeling the test score gains of repeated cross sections of students. As discussed in more detail below, this model is likely to produce similar value-added estimates as the commonly used Education Value-Added Assessment System (EVAAS) model (Sanders, Saxton, & Horn, 1997), where longitudinal data are used to directly model the growth in a student's test scores over time.

Finally, the above analysis assumes that gain scores are used from a single academic subject only. However, value-added estimates for upper elementary school teachers are sometimes obtained using test scores from *multiple* subject areas. Our primary analysis assumes a test score from a single subject (or from highly correlated tests), but in our sensitivity analysis, we examine precision gains from using multiple tests (as discussed further below).

Considered Estimators

Although maximum likelihood estimators based on the expectation-maximization (EM) algorithm are typically used to estimate HLMs, the same estimators can be obtained through simpler methods due to the balanced design assumed here. We consider two estimators for τ_{jk} using variants of the HLM in (1a) to (1d). The first is an OLS estimator that is obtained using the following model, where (1b) is inserted into (1a) and τ_{jk} are treated as fixed effects:

$$g_{itjk} = \tau_{jk} + (\omega_{tjk} + \varepsilon_{itjk}). \tag{2}$$

This model yields the following OLS estimator:

$$\hat{\tau}_{jk,\mathrm{OLS}} = \bar{g}_{..jk}, \tag{3}$$

where $\bar{g}_{..jk} = \left(\sum_{t=1}^{c}\sum_{i=1}^{n} g_{itjk}/cn\right)$ is the mean gain score for all students taught by teacher j in school k over the c years.

The second approach for estimating τ_{jk} is an EB approach (see, e.g., Berger, 1985; Lindley & Smith, 1972; Raudenbush & Bryk, 2002). The EB estimator for τ_{jk} is the mean of the posterior distribution for τ_{jk} given the data.

There are several EB estimators that could be used for between-school comparisons. One estimator, which involves application of the EB approach twice, uses the four-level HLM from above and is as follows:

$$\hat{\tau}_{jk,\mathrm{EB,Between1}} = \lambda_\theta \bar{g}_{..jk} + (1-\lambda_\theta)[\lambda_\psi \bar{\bar{g}}_{...k} + (1-\lambda_\psi)\bar{\bar{\bar{g}}}_{....}], \tag{4}$$

where $\bar{\bar{\bar{g}}}_{....} = \hat{\delta} = \left(\sum_{k=1}^{s} \bar{\bar{g}}_{...k}/s\right)$ is the grand district-level mean, $\lambda_\theta = \sigma^2_\theta/(\sigma^2_\theta + \sigma^2_{\bar{g}|\tau})$ is the within-school "reliability" weight $(0 \leq \lambda_\theta \leq 1)$, $\sigma^2_{\bar{g}|\tau} = (\sigma^2_\omega/c) + (\sigma^2_\varepsilon/cn)$ is the variance of $\bar{g}_{..jk}$ conditional on τ_{jk}, $\lambda_\psi = \sigma^2_\psi/(\sigma^2_\psi + \sigma^2_{\bar{\bar{g}}|\eta})$ is the between-school reliability weight $(0 \leq \lambda_\psi \leq 1)$, and $\sigma^2_{\bar{\bar{g}}|\eta} = (\sigma^2_\theta/m) + (\sigma^2_\omega/cm) + (\sigma^2_\varepsilon/cnm)$ is the variance of $\bar{\bar{g}}_{...k}$ conditional on η_k. This estimator "shrinks" the teacher-level mean toward the school-level mean, which, in turn, is shrunk toward the grand district-level mean (Raudenbush & Bryk, 2002).

The EB estimator in Equation (4) could lead to the result that a teacher with a higher value for $\bar{g}_{..jk}$ than another teacher is given a lower performance ranking because she teaches in a lower-performing school (which may be difficult to explain to educators). Thus, we also consider a second EB estimator that does not adjust for performance differences across schools. It can be obtained from a three-level HLM where (1d) is inserted into (1c). This estimator is as follows:

$$\hat{\tau}_{jk,\mathrm{EB,Between2}} = \lambda_{\tau'} \bar{g}_{..jk} + (1-\lambda_{\tau'})\bar{\bar{\bar{g}}}_{....}, \tag{5}$$

where $\lambda_{\tau'} = (\sigma^2_\theta + \sigma^2_\psi)/(\sigma^2_\theta + \sigma^2_\psi + \sigma^2_{\bar{g}|\tau})$ is the between-school reliability weight $(0 \leq \lambda_{\tau'} \leq 1)$. This estimator directly shrinks the teacher-level mean to the grand district-level mean.

For several reasons related to the clarity of presentation, we adopt the estimator in (5) rather than (4) for our analysis. First, for the system error rate analysis, we must calculate the variance of the EB estimator, and the variance formula is very complex for the estimator in Equation (4) (see Berger, 1985). Second, it is more straightforward to use Equation (5) rather than Equation (4) to specify null hypotheses for comparing τ_{jk} values for teachers across different schools so that the test statistics (z scores) have zero expectations under the null hypotheses.

Third, the estimator in Equation (5) is used most prevalently in practice (Lipscomb, Teh, Gill, Chiang, & Owens, 2010).

A Representative Scheme for Comparing Teacher Performance

In any performance measurement system, there must be a decision rule for classifying teachers as meriting or not meriting special treatment. One of the most prevalent value-added models applied in practice is the EVAAS model used by the Teacher Advancement Program (TAP; see National Institute for Excellence in Teaching, 2009), which classifies each teacher into a performance category based on the *t* statistic from testing the null hypothesis that the teacher's performance is equal to the average performance in a reference group (see Solmon, White, Cohen, & Woo, 2007; Springer, Ballou, & Peng, 2008). Thus, hypothesis testing is an integral part of the policy landscape in performance measurement, and forms the basis for our considered scheme for comparing teacher value-added estimates. We emphasize, then, that the error rate formulas and empirical calculations of this article pertain only to performance measurement systems that use hypothesis testing to classify educator performance. This approach does not encompass several other types of decision rules that could be used, such as identifying teachers whose value-added estimates are simply above or below fixed thresholds or percentiles of the teacher quality distribution, which ignores the variance of the estimates.

In particular, this section considers a performance measurement scheme that addresses the question, "Which teachers performed particularly well or badly relative to the average teacher in the school district?" The scheme assumes a classical hypothesis testing strategy for both the OLS and EB estimators. Under this scheme, the considered null hypothesis is $H_0 : \tau_{jk} - \bar{\bar{\tau}}_{..} = 0$, where $\bar{\bar{\tau}}_{..} = \left(\sum_{k=1}^{s} \bar{\tau}_{.k}/s \right)$ is the mean value of τ_{jk} across all teachers in the district. This testing approach will identify for special treatment teachers for whom the null hypothesis is rejected.[2]

Using the *OLS approach* and the estimator in Equation (3), the null hypothesis for our performance scheme can be tested using the *z* score $z_{\mathrm{OLS}} = [(\bar{g}_{..jk} - \bar{\bar{\bar{g}}}_{....})/\sqrt{V_{\mathrm{OLS}}}]$, where the variance V_{OLS} is defined as follows:

$$V_{\mathrm{OLS}} = \left(\frac{\sigma_{\omega}^2}{c} + \frac{\sigma_{\varepsilon}^2}{cn} \right) \left(\frac{sm-1}{sm} \right). \qquad (6)$$

For moderate or large *m*, the variance in Equation (6) is driven primarily by the variance of the teacher-level mean (because the second bracketed term is close to 1). Thus, our considered scheme has similar statistical properties to a more general scheme where statistical tests are conducted to compare a teacher's performance relative to fixed threshold values that are assumed to be measured without error.

Using the EB approach and the estimator in Equation (5), it is still conventional to use classical test statistics to test the null hypothesis (Raudenbush & Bryk, 2002, chap. 3). For our performance scheme, the null hypothesis can be tested with the z score $z_{EB} = [\lambda_{\tau'}(\bar{g}_{..jk} - \bar{\bar{\bar{g}}}_{....})]/\sqrt{V_{\text{EB}}}$, where V_{EB} is an approximation to the variance of the posterior distribution for τ_{jk} given the data. This approximation, which assumes that $\hat{\eta}_k = \bar{\bar{g}}_{...k}$ and $\hat{\delta} = \bar{\bar{\bar{g}}}_{....}$ are estimated without error, is defined as follows:

$$V_{\text{EB}} = \lambda_{\tau'} V_{\text{OLS}}\left(\frac{sm}{sm-1}\right) = \lambda_{\tau'}\sigma^2_{\bar{g}|\tau}, \tag{7}$$

where $\lambda_{\tau'}$ is the reliability weight defined above (see Berger, 1985; Gelman, Hill, & Yajima, 2009).

The variance of the EB estimator in Equation (7) is smaller than the variance of the OLS estimator in Equation (6) by a factor of $\lambda_{\tau'} sm/(sm-1)$. However, the z score is *smaller* for the EB estimator by a factor of $\sqrt{\lambda_{\tau'}(sm-1)/sm}$, suggesting that the OLS approach tends to reject the null hypotheses more often—and thus has *more* statistical power—than the EB approach. This occurs because as $\lambda_{\tau'}$ decreases, the teacher-level means are shrunk toward the district-level mean faster than the EB standard errors are shrunk toward zero.

Finally, we assume one-sided rather than two-sided tests, where the direction of the alternative hypothesis depends on the sign of the z score. For instance, if a teacher's value-added estimate is observed to be below average, it is likely that district officials will then want to test whether the teacher's *true* performance is below average, which naturally serves as the alternative hypothesis for this test.

Accounting for Tests From Multiple Subjects

The above analysis assumes that value-added estimates are obtained using gain scores from a single academic subject. However, we also consider the case in which teacher effects are estimated separately for each subject area, and are then appropriately scaled and averaged to obtain aggregate value-added estimates of a teacher's underlying τ_{jk} value.

Suppose that gain score data are available for d subject tests per student, so that nd test score observations are available for each classroom. These d test score observations are "clustered" within students. Standard methods to calculate the variance of two-stage clustered designs (see, e.g., Kish, 1965) can then be used to show that the *effective* number of observations per classroom can be approximated as follows:

$$n_{\text{eff}} \approx nd/[1 + \rho_d(d-1)], \tag{8}$$

where ρ_d is the average pairwise correlation among the d tests. The denominator in Equation (8) is the design effect and will increase as the correlations between the subject tests increase.

To account for multiple tests in the above variance formulas, the effective class size n_{eff} can be used rather than n. Realistic values for d and ρ_d are discussed below.

Accounting for the Serial Correlation of Student Gain Scores Over Time

The benchmark HLM uses only contemporaneous information on student test score gains to estimate the value-added of the students' current teachers. However, some models such as EVAAS pool together gain scores from multiple grades for a given student (see Schochet & Chiang, 2010, for a more detailed discussion). Because the student-level error term, ε_{itjk}, may be serially correlated across grades, the gains achieved by a teacher's *current* students in *other* grades reduce uncertainty about the students' current values of ε_{itjk} and, hence, can improve the precision of the value-added estimates.

Suppose that the HLM in (1a) through (1d) is specified separately for two consecutive grades, and the two grade-specific models are estimated jointly as a system of seemingly unrelated regressions (SURs) using generalized least squares (GLS; see, e.g., Amemiya, 1985). By comparing well-known variance formulas for the GLS estimator and the OLS estimator (Wooldridge, 2002, chap. 7), it can be shown that the variance of the SUR estimator is approximately the variance of the OLS estimator based on the effective number of observations per classroom, $n_{\text{eff2}} \approx n/(1-\rho_{t,t-1}^2)$, where $\rho_{t,t-1}$ is the correlation of ε_{itjk} across the two grades. In our sensitivity analysis, we present results using n_{eff2} rather than n, using empirical values of $\rho_{t,t-1}$. As discussed below, empirical values of error correlations across nonconsecutive grades are very small, so we do not consider the case of pooling more than two grades together.

Accounting for Additional Precision Gains From Controlling for Pretest Scores

Student test score gains could vary depending on pretest score values. Thus, additional precision gains could be realized by including pretest scores as a covariate in the HLM. The variance formulas can be modified to reflect the presence of this covariate. For the OLS estimator, the variance formula in Equation (6) reflects this covariate adjustment if the error variances, σ_ω^2 and σ_ε^2, are multiplied respectively by $(1-R_\omega^2)$ and $(1-R_\varepsilon^2)$, where R_ω^2 and R_ε^2 are the proportions of the classroom-level and student-level error variance explained by pretest scores. A similar adjustment can be applied to the variance of the EB estimator. In our sensitivity analyses, we use empirically based R-square values and apply these modified formulas.

Measuring the Reliability of Performance Estimators

The variance formulas presented above have a direct relation to reliability (stability) measures that the previous literature has used to gauge the noisiness in value-added estimators (see, e.g., McCaffrey et al., 2009). Parallel to its usual psychometric definition, reliability in this context is the proportion of an estimator's total variance across teachers that is attributable to the "signal"—that is, persistent performance differences across teachers. In our notation, the reliability of the OLS estimator is $\lambda_{\tau'}$ (as defined in Equation (5) above). We report reliability statistics in the empirical work below, but this metric is not entirely consistent with our hypothesis testing framework. Thus, as discussed next, our focus is on false positive and negative error rates to measure the accuracy of a performance measurement system based on hypothesis testing.

Calculating System Error Rates

In this section, we discuss our approach for defining system error rates under our considered testing scheme, and key issues that must be considered when applying these definitions.

Defining System Error Rates

We define system error rates using false positive and negative error rates from classical hypothesis testing. To help explain our error rates, consider our scenario where a hypothesis test is conducted to assess whether a teacher performs significantly worse than the average teacher in her district using test score data for *c* years. Suppose also that a teacher is considered to be in need of special treatment if her true performance level is *T* gain score standard deviations (*SD*s) below the district average (see Figure 1).[3] We assume this scenario for the remainder of this section; symmetric results apply for tests that aim to identify high-performing teachers.

The Type I error rate (α) is the probability that based on *c* years of data, the hypothesis test will find that a truly average teacher (such as Teacher 4) performed significantly worse than average. Given α, the false positive error rate, $\text{FPR}(q)$, is the probability that a teacher (such as Teacher 5) whose true performance level is *q SD*s above average is falsely identified for special treatment. For a one-tailed *z* score test, $\text{FPR}(q)$ can be expressed as follows:

$$\text{FPR}(q) = \Pr(\text{Reject } H_0 | \tau_{jk} - \bar{\bar{\tau}}_{..} = q\sigma) = 1 - \Phi\ [\Phi^{-1}(1-\alpha) + \frac{q\sigma\lambda}{\sqrt{V}}]\ \textit{for } q \geq 0, \quad (9)$$

where $\sigma^2 = (\sigma^2_\psi + \sigma^2_\theta + \sigma^2_\omega + \sigma^2_\varepsilon)$ is the *total* variance of the student gain score, V is the variance of the OLS or EB estimator, λ equals 1 for the OLS estimator and $\lambda_{\tau'}$ for the EB estimator, and $\Phi(.)$ is the normal distribution function.

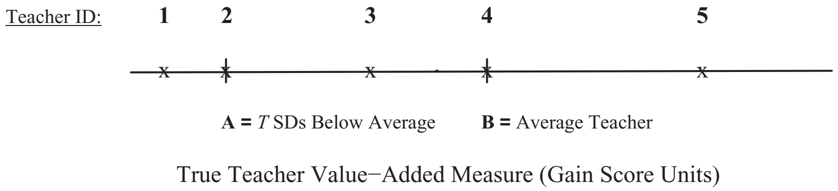

FIGURE 1. *Hypothetical true teacher value-added measures.*

Equation (9) makes clear that at any given $q > 0$, the EB estimator has a larger false positive error rate than the OLS estimator because $(\lambda/\sqrt{V})$ is smaller for the EB estimator. For teachers who are truly above average, shrinkage causes the distribution of the EB estimator to be centered at a lower value than the OLS estimator, implying a greater probability that the EB estimator incorrectly identifies these teachers as low performing.

The *overall* false positive error rate for the population of average or better teachers can be obtained by calculating the expected value (weighted average) of population $\text{FPR}(q)$ values:

$$\text{FPR\_TOT} = \int\limits_{q\geq 0} \Pr(\text{Reject } H_0 | \tau_{jk} - \bar{\bar{\tau}}_{..} = q) f(q | \tau_{jk} - \bar{\bar{\tau}}_{..} \geq 0) \partial q, \qquad (10)$$

where $f(.)$ is the density of q for the considered population. Clearly, $\text{FPR\_TOT} \leq \alpha$. For the empirical analysis, we assume that q has a normal distribution with variance $V_f = (\sigma_\psi^2 + \sigma_\theta^2)/\sigma^2$. To calculate the integral in Equation (10), we used a simulation estimator where we obtained 10,000 random draws for q from a truncated normal distribution, calculated $\text{FPR}(q)$ for each draw, and averaged these 10,000 false positive error rates.

Given α and the threshold value T, the *false negative error rate* is the probability that the hypothesis test will fail to identify teachers (such as Teachers 1 and 2 in Figure 1) whose true performance is at least T *SD*s below average. For a one-tailed z score test, $\text{FNR}(q)$ can be expressed as follows:

$$\text{FNR}(q) = \Pr(\text{Do Not Reject } H_0 | \tau_{jk} - \bar{\bar{\tau}}_{..} = q\sigma) = \Phi\left[\Phi^{-1}(1-\alpha) + \frac{q\sigma\lambda}{\sqrt{V}}\right] \qquad (11)$$

for $q \leq T < 0$. The Type II error rate, $(1-\beta)$, equals $\text{FNR}(T)$, where β is the statistical power level. For every $q \leq T < 0$, the EB estimator has a higher false negative error rate than the OLS estimator.

The *overall* false negative error rate for the population of low-performing teachers can be calculated as follows:

$$\mathrm{FNR\_TOT} = \int\limits_{q \le T < 0} \mathrm{FNR}(q) f(q | \tau_{jk} - \bar{\bar{\tau}}_{..} \le T\sigma) \partial q. \tag{12}$$

Note that in Equation (12) we do not include teachers whose performance values are *between* points A and B in Figure 1 (such as Teacher 3). This is because it is difficult to assess whether failure to identify these teachers should be regarded as an error or not. Although these teachers do not truly perform poorly enough for the district to regard them as needing special treatment, their true performance lies in the range covered by the alternative hypothesis of the statistical test. Because of this ambiguity, we exclude these teachers from the error rate calculations, but, in our empirical analyses, we conduct calculations assuming different threshold values.

We also analyze two additional aggregate error rates—first discussed by Benjamini and Hochberg (1995)—that have a Bayesian interpretation. First, for a given α, we define the population *false discovery rate*, FDR_TOT, as the expected proportion of all teachers with significant test statistics who are false discoveries (i.e., who are truly average or better). Using Bayes rule, FDR_TOT can be approximated as follows:

$$\mathrm{FDR\_TOT} \approx \frac{\mathrm{FPR\_TOT}^{*}.5}{\int\limits_{q} \Pr(\text{Reject } H_0 | \tau_{jk} - \bar{\bar{\tau}}_{..} = q\sigma) f(q) \partial q}. \tag{13}$$

Second, for a given α and T, we define the *false non-discovery rate*, FNDR_TOT, as the expected proportion of all teachers with insignificant test statistics who are truly low performers. This error rate can be approximated as follows:

$$\mathrm{FNDR\_TOT} \approx \frac{\mathrm{FNR\_TOT} \times (1 - \Phi(|T|\sigma/\sqrt{V_f})}{\int\limits_{q} \Pr(\text{Do Not Reject } H_0 | \tau_{jk} - \bar{\bar{\tau}}_{..} = q\sigma) f(q) \partial q}. \tag{14}$$

Tolerable error rates will likely depend on a number of factors (see Schochet & Chiang, 2010, for a more extensive discussion). First, they are likely to depend on the nature of the system's rewards and penalties. For instance, acceptable levels are likely to be lower if the system is to be used for making high-stakes decisions. Second, acceptable error rate levels are likely to differ by stakeholder (such as teachers, students, parents, school administrators, and policymakers) who may have different loss functions for weighing the various benefits and costs of correct and incorrect system classifications. Because it is not possible to define universally acceptable levels for system error rates, our empirical analysis presents error rates for different assumed numbers of years of available data and is agnostic about what error rate levels are appropriate and what types of errors are more serious.

We consider Type I and II error rates as well as the overall error rate measures discussed above. Type I and II error rates provide an *upper bound* on system error rates for individual teachers or schools. These rates may be applicable if the system is to be used for making high-stakes decisions and stakeholders view misclassification errors as highly consequential. The Type I error rate may also be of particular interest to teachers or schools who believe that their performance is at least average (but who are not sure how much above average), because this rate is the maximum chance that such a teacher or school will be falsely identified for sanctions. The Type II error rate may be of particular interest to administrators and parents who want a conservative estimate of the chances that a very low-performing teacher will be missed for special treatment and remain in the classroom without further intervention.

We also present results using the *overall* error rate measures for those interested in aggregate misclassification rates for the full population of "good" and "poor" educators. Such interested parties might include designers of accountability systems whose focus is on the social equity of a performance measurement system. Thus, the Type I and II error rates may be relevant to those focused on individual teachers and schools, whereas the overall error rates may be more relevant to those focused on groups of educators.

In order to balance the myriad objectives from above and keep the presentation of empirical results manageable, we report results from three types of analyses. First, we report Type I and II error rates subject to the restriction that these two error rates are equal, consistent with an approach of being agnostic about which type of error is more serious. For given values of c and T, these error rates can be calculated as follows:

$$\alpha = 1 - \beta = 1 - \Phi\left[\frac{|T|\sigma\lambda}{2\sqrt{\text{Var}(\text{Contrast})}}\right], \tag{15}$$

where Var(Contrast) is the variance of the contrast of interest and other terms are defined as above. Second, we use a grid search using different Type I errors to calculate and report values for FPR_TOT and FNR_TOT subject to the restriction that these two error rates are equal. For these derived values, we also present FDR_TOT and FNDR_TOT values. Finally, because some stakeholders may place different weights on false negatives and positives, we report results on the number of years of available data per teacher that are required to attain various combinations of Type I and II errors and FPR_TOT and FNR_TOT errors.

Statistical Framework for the School-Level Analysis

The statistical framework from above can also be used to identify *schools* for special treatment. These methods can be implemented using estimates of $\eta_k = (\delta + \psi_k)$ from variants of the HLM from above. For the OLS approach, estimates of η_k can be obtained using the following model, where η_k and θ_{jk} are treated as fixed effects:

$$g_{itjk} = \eta_k + \theta_{jk} + (\omega_{tjk} + \varepsilon_{itjk}). \tag{16}$$

The resulting OLS estimator is $\hat{\eta}_{k,\text{OLS}} = \bar{\bar{g}}_{...k}$.

The EB estimator for η_k can be obtained using the four-level HLM and is $\hat{\eta}_{k,EB} = \lambda_\psi \bar{\bar{g}}_{...k} + (1 - \lambda_\psi)\bar{\bar{\bar{g}}}_{....}$, where λ_ψ is the between-school reliability weight defined above. Unlike the OLS framework, the EB framework treats η_k and θ_{jk} as random rather than fixed.

For the school-level analysis, the null hypothesis for comparing a school's performance to the district average is $H_0: \psi_k - \bar{\psi}_. = 0$, where $\bar{\psi}_. = (\sum_{k=1}^{s} \psi_k / s)$ is the mean school effect in the district. Using the *OLS approach,* this null hypothesis can be tested using the z score $z_{\text{OLS,School}} = [(\bar{\bar{g}}_{...k} - \bar{\bar{\bar{g}}}_{....})/\sqrt{V_{\text{OLS,School}}}]$, where $V_{\text{OLS,School}}$ is defined as follows:

$$V_{\text{OLS,School}} = \left(\frac{\sigma^2_\omega}{cm} + \frac{\sigma^2_\varepsilon}{cnm}\right)\left(\frac{s-1}{s}\right). \tag{17}$$

The z score using the *EB approach* is as follows:

$$z_{2,\text{EB,School}} = \left[\lambda_\psi(\bar{\bar{g}}_{...k} - \bar{\bar{\bar{g}}}_{....})/\sqrt{\lambda_\psi \sigma^2_{\bar{\bar{g}}|\eta}}\right]. \tag{18}$$

Our approach for assessing appropriate sample sizes for the school-based analysis is parallel to the approach discussed above for the teacher-based analysis.

Empirical Analysis

In this section, we calculate system error rates for the performance measurement schemes and estimators considered above, using empirically based values for key parameters.

Obtaining Realistic Values for the Variance Components

The error rate formulas from above depend critically on the variances of the specific performance contrasts. These variances are functions of the intraclass correlations (ICCs) $\rho_\psi = \sigma^2_\psi/\sigma^2$, $\rho_\theta = \sigma^2_\theta/\sigma^2$, $\rho_\omega = \sigma^2_\omega/\sigma^2$, and $\rho_\varepsilon = \sigma^2_\varepsilon/\sigma^2$, which express the variance components in (1a) to (1d) as fractions of the total variance in gain scores across all students in the district.

In a companion report, Schochet and Chiang (2010) obtained realistic ICC estimates by reviewing 10 recent studies from the value-added literature that provide information on at least one ICC, and by conducting primary analyses of data from five large-scale experimental evaluations of elementary school interventions.[4] The average ICCs across these studies are the benchmark ICCs that were used in the analyses.

These analyses show that student heterogeneity is the key source of imprecision in estimating differences in value-added across teachers and schools. On average, 92% of the total gain score variance is attributable to student differences within the same classroom ($\rho_\varepsilon = 0.92$). Another source of imprecision stems from idiosyncratic classroom-level factors, which, on average, account for 3% of the total variance in gain scores ($\rho_\omega = 0.030$). In addition, the proportion of the total variance that is attributable to persistent, within-school differences in teacher value-added is about 3.5% ($\rho_\theta = 0.035$). School-level factors account for an additional 1.1% of the gain score variance ($\rho_\psi = 0.011$).

Additional Assumptions for Key Parameters

Other parameters that enter the error rate formulas are the class size (n), the number of teachers per school (m), and the number of schools in the district (s). We assume $n = 21$, which is the median class size for self-contained classrooms in elementary schools according to our calculations from the 2003-2004 School and Staffing Survey (SASS).

The assumed value of m depends on the number of elementary grade levels that are likely to be included in a performance measurement scheme. Under No Child Left Behind (NCLB), state assessments must begin no later than third grade, but some states and districts administer assessments to earlier grades. Therefore, we make the assumption that each elementary school has three grade levels for which teacher value-added can be estimated, and that these three grades collectively have $m = 10$ teachers (or an average of 3.3 teachers per grade). This assumption yields 70 students per grade level per school, which is approximately the median fourth grade enrollment of elementary schools in 2006–2007 according to our calculations from the Common Core of Data.

We assume multiple values of s because districts vary widely in size. In particular, we present results for $s = 5$ and $s = 30$, which imply districtwide fourth grade enrollment in the 81st and 98th percentiles. Our focus on the top quintile of district size stems from the fact that districts in this quintile educate more than 70% of the nation's students.

For our sensitivity analysis, we require values of d (the number of tests) and ρ_d (the average pairwise correlation between student gain scores from multiple tests). For only two subjects, reading and math, does NCLB mandate assessments in consecutive grades. Thus, we assume $d = 2$ for the sensitivity analysis. To obtain realistic values for ρ_d, we calculated correlations between math and reading fall-to-spring gain scores using the experimental data discussed above. These correlations range from .2 to .4. Thus, for our analysis, we assume $\rho_d = 0.3$. Applying (8) with $d = 2$, $\rho_d = 0.3$, and $n = 21$ yields an effective sample size, n_{eff}, equal to 32 students per classroom.

The sensitivity analysis that exploits longitudinal student information from consecutive grades requires values of $\rho_{t,t-1}$, the correlation of ε_{itjk} across consecutive grades. Using data on the population of North Carolina's students in third, fourth, and fifth grades, Rothstein (2010) finds that the correlation in gain score residuals between fourth and fifth grade is $-.38$ in math and $-.37$ in reading. Thus, our sensitivity analysis uses $\rho_{t,t-1} = -0.38$, implying an effective sample size of $n_{\text{eff},2} = 24.5$ students per classroom. Rothstein also finds that the correlation between gain scores in grades three and five ranges from $-.02$ to .02; thus, we ignore error correlations across nonconsecutive grade levels.

The sensitivity analysis also requires values for R^2_ω and R^2_ε, the *R*-square at the classroom and student levels from controlling for pretest scores. In the experimental data discussed above, the average value of R^2_ε was .17. Because data from each study covered only 1 year, classroom-level variance components could not be distinguished from teacher-level variance components, so it was not feasible to obtain empirical values of R^2_ω. For the sensitivity analysis, we assume that $R^2_\omega = R^2_\varepsilon = 0.17$.

Identifying Threshold Values

A critical issue is the threshold to adopt for defining meaningful performance differences between teachers or schools (that is, the value of *T* in Figure 1). Following the approach used elsewhere (Bloom, Hill, Black, & Lipsey, 2008; Kane, 2004; Schochet, 2008), we identify educationally meaningful thresholds using the natural progression of student test scores over time.

To implement this approach, we use estimates of average annual gain scores compiled by Bloom et al. (2008). Because their estimates are expressed in posttest score *SD*s, we converted them to *SD*s of gain scores by dividing them by .696, the estimated ratio of the *SD* of test score gains to the *SD* of posttest scores from the experimental data discussed above. On average, annual growth in achievement per grade is .65 *SD*s of reading gains and .94 *SD*s of math gains. Based on these estimates, we conduct our calculations for the teacher analyses using threshold values of .1, .2, and .3 *SD*s. A .2 value represents 31% of an average annual gain score in reading, or about 4 months of reading growth attained by a typical upper elementary student; in math, it represents 21% of an average annual gain score, or about 3 months of student learning. These differences are large relative to the distribution of true teacher value-added: using our ICC estimates, an increment of .2 *SD*s in student gain scores is equivalent to the performance difference between a district's 50th percentile and 82nd percentile teachers, which is consistent with findings from previous literature (Hanushek & Rivkin, 2010).

We use smaller threshold targets for the school analysis than for the teacher analysis since variation in school value-added is smaller than the within-school variation in teacher value-added (that is, ρ_ψ values tend to be smaller than ρ_θ values). For school comparisons, we use thresholds that are half the size of those from the teacher comparisons; the resulting values of .05, .1, and .15 represent the differences between a district's 50th percentile school and, respectively, its 68th percentile, 83rd percentile, and 92nd percentile schools.

Empirical Results

Tables 1 through 7 provide the main empirical findings. The key results can be summarized as follows:

Finding 1: Using sample sizes typically available in practice, the considered performance measurement systems for teacher-level analyses will generally yield Type I and II error rates of at least 20%. Consider a system that aims to identify high-performing teachers in the upper elementary grades using sample sizes typically available in practice (1–5 years of data per teacher). Suppose also that policymakers find it acceptable to set $\alpha = 1 - \beta$ and to set the threshold level for defining a high-performing teacher at .2 *SD*s above the average performance level. In this case, with $c = 3$ years of data, a scheme that compares a teacher to the district average would yield a Type I or II error rate of 26% using the OLS estimator (Table 1). In other words, the system would miss for recognition one fourth of truly high-performing teachers who are at the 82nd percentile of performance in their district, and would erroneously identify for recognition one fourth of persistently average teachers. The error rates are about 2–5 percentage points larger for the EB than OLS estimator (Table 1).

Type I and II error rates would exceed one third with only 1 year of data and would drop to one fifth with 5 years of data (Table 1). The error rates would increase by about 10 percentage points using a threshold value of .1 *SD*s rather than .2 *SD*s, and would decrease by about 10 percentage points using a threshold value of .3 *SD*s (Table 1). Parallel results apply to systems that aim to identify below-average performers.

Consistent with error rates dropping as a function of c, reliability of the considered estimators rises with c. The reliability of the OLS estimator (as measured using $\lambda_{\tau'}$ in Equation 5 above) is .38 for $c = 1$, .65 for $c = 3$, .76 for $c = 5$, and .86 for $c = 10$.

The Type I error pertains to teachers of average performance, and the Type II error pertains to teachers whose performance is at the selected threshold value. Misclassification rates for teachers as a group, however, are better captured by overall false positive and negative error rates. This consideration motivates our second key finding that is discussed next.

Finding 2: Overall false negative and positive error rates for identifying low- and high-performing teachers are likely to be smaller than Type I and II error rates. Suppose that FNR_TOT and FPR_TOT are restricted to be equal, and

TABLE 1
Teacher-Level Analysis: Type I and II Error Rates That Are Restricted to Be Equal, by Threshold Value, Scheme, and Estimator

	Threshold Value (Gain Score *SD*s From the Average)[a]					
	OLS			Empirical Bayes (EB)		
Number of years of available data per teacher	.1	.2	.3	.1	.2	.3
Compare a teacher to the district average (50 teachers in the district)						
1	.43	.36	.29	.45	.41	.37
3	.37	.26	.17	.40	.30	.22
5	.34	.20	.11	.36	.24	.14
10	.28	.12	.04	.29	.14	.05
Compare a teacher to the district average (300 teachers in the district)						
1	.43	.36	.29	.45	.41	.37
3	.37	.26	.17	.40	.30	.22
5	.34	.20	.11	.36	.24	.14
10	.28	.12	.04	.29	.14	.05

Note. OLS = ordinary least squares. See the text for formulas and assumptions. Calculations assume test score data from a single subject area.
[a]See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

assume a threshold value of .2 *SD*s and $c = 3$. In this case, FNR_TOT and FPR_TOT equal 10% for the OLS estimator (Table 2), whereas α and $(1-\beta)$ equal 26% when equated (Table 1). The corresponding error rates using the EB estimator are 14% and 30%, respectively. A similar pattern holds for other threshold values

Consistent with these findings, fewer numbers of years of data are required to achieve various combinations of overall error rates than the corresponding combinations of Type I and II error rates (Table 3). For example, using a threshold value of .2 *SD*s, the OLS estimator would require about 3 years of data to ensure that FPR_TOT = 0.05 and FNR_TOT = 0.20, compared to 11 years to ensure that $\alpha = .05$ and $(1 - \beta) = 0.20$ (Table 3).

Finally, Table 4 shows false discovery and nondiscovery rates using the FPR_TOT and FNR_TOT values from Table 2. Assuming $c = 3$ and a threshold value of .2 *SD*s, the OLS estimator yields an FDR_TOT value of 13%. This means that slightly more than one eighth of teachers who are identified for special treatment are expected to be false discoveries. For this same scenario, FNDR_TOT is 3%, which means that only a small percentage of all teachers with insignificant test statistics are expected to be misclassified.

TABLE 2
Teacher-Level Analysis: Overall False Positive and Negative Error Rates That Are Restricted to Be Equal

	Threshold Value (Gain Score *SD*s From the Average)[a]					
	OLS			Empirical Bayes (EB)		
Number of years of available data per teacher	.1	.2	.3	.1	.2	.3
Compare a teacher to the district average (50 teachers in the district)						
1	.24	.20	.16	.33	.30	.27
3	.15	.10	.06	.19	.14	.10
5	.10	.06	.03	.13	.08	.05
10	.06	.02	.01	.07	.03	.01
Compare a teacher to the district average (300 teachers in the district)						
1	.25	.21	.17	.33	.30	.27
3	.15	.10	.06	.19	.14	.10
5	.11	.06	.03	.13	.08	.05
10	.06	.02	.01	.07	.03	.01

Note. OLS = ordinary least squares. See the text for formulas and assumptions. Calculations assume test score data from a single subject area.
[a] See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

Finding 3: The empirical results for the teacher-level analysis are robust to alternative ICC assumptions, the use of two subject tests, the use of two successive years of gain scores on each student, and the inclusion of controls for pretest scores. Our benchmark calculations were conducted using the average of the ICC values reported in Schochet and Chiang (2010). However, because of the uncertainty in the ICC estimates and their critical role in the calculations, we examined the sensitivity of our main results by varying the key parameters ρ_ω and ρ_ε. As shown in Table 5, even if ρ_ω and ρ_ε were much lower than the benchmark values, the error rates would, in general, remain similar to those in Tables 1 and 2. Reducing ρ_ε from .92 (the baseline assumption) to .80 (a value lower than all but one estimate in Schochet & Chiang, 2010), or reducing ρ_ω to zero (which is equivalent to assuming that classroom effects are fixed and reflect a teacher's true performance in a given year), leaves the Type I and II error rates at a minimum of 20% (assuming $c = 3$). The error rates for FPR_TOT and FNR_TOT, however, are somewhat more sensitive to lowering or raising the ICC values than the Type I and II errors.

Allowing for multiple tests (e.g., math and reading) instead of one test has little effect on the error rate estimates (Table 5). For instance, with $c = 3$, allowing

TABLE 3

Teacher-Level Analysis: The Number of Years of Data Required to Achieve Various System Error Rates

	Type II Error Rate/Overall False Negative Rate			
Type I Error Rate/Overall False Positive Rate	.05	.10	.15	.20
Threshold value = .1 *SD*s[a]				
.05	78/12	62/8	52/6	45/5
.10	62/8	48/5	39/4	33/3
.15	52/6	39/4	31/3	26/2
.20	45/5	33/3	26/2	20/2
Threshold value = .2 *SD*s[a]				
.05	20/6	15/4	13/3	11/3
.10	15/4	12/3	10/2	8/2
.15	13/3	10/2	8/2	6/1
.20	11/3	8/2	6/1	5/1
Threshold value = .3 *SD*s[a]				
.05	9/4	7/3	6/2	5/2
.10	7/3	5/2	4/2	4/1
.15	6/2	4/2	3/1	3/1
.20	5/2	4/1	3/1	2/1

Note: In cells with two entries, the first entry represents the number of years required to achieve Type I and Type II error rates represented, respectively, by the row and column headers; the second entry represents the number of years required to achieve overall false positive and false negative rates represented, respectively, by the row and column headers. Figures are based on the ordinary least squares (OLS) estimator and assume test score data from a single subject area. The figures correspond to a scheme where a teacher is compared to the district average with 50 teachers in the district. See the text for formulas and assumptions.

[a] See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

for two tests decreases the Type I or II error rate from .26 to .24 and the overall false positive or negative rate from .10 to .08.

Likewise, there are negligible reductions in error rates from using students' gain scores in current and adjacent years rather than in the current year only (Table 5). For $c = 3$, both the Type I or II error rates and the overall false positive or negative error rates would decline by only 1 percentage point using the longitudinal approach. These analyses suggest that our benchmark findings, which are based on contemporaneous gain score data only, are likely to be applicable to value-added models such as EVAAS that exploit longitudinal gain score data on each student.

Error rates decrease only slightly when pretest scores are included as a covariate in the HLM (Table 5). Adding this covariate leads to a decrease in error rates

TABLE 4
Teacher-Level Analysis: False Discovery and Nondiscovery Rates Using the Overall False Positive and Negative Error Rates in Table 2

	Threshold Value (Gain Score *SD*s From the Average)[a]					
	0.1		0.2		0.3	
Number of Years of Available Data Per Teacher	False Discovery Rate	False Nondiscovery Rate	False Discovery Rate	False Nondiscovery Rate	False Discovery Rate	False Nondiscovery Rate
1	.27	.14	.25	.06	.23	.02
3	.17	.08	.13	.03	.10	.01
5	.12	.06	.08	.02	.05	.00
10	.07	.03	.03	.01	.01	.00

Note. Figures are based on the OLS estimator and assume test score data from a single subject area. The figures correspond to a scheme where a teacher is compared to the district average with 50 teachers in the district. See the text for formulas and assumptions.

[a] See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

TABLE 5

Teacher-Level Analysis: Sensitivity of System Error Rates to Key ICC Assumptions and the Use of Multiple Tests

	Type I = Type II Error Rate/False Negative = False Positive Error Rate: For a Threshold Value of .2 *SDs*[a]					
	Number of Years of Data = 1		Number of Years of Data = 3		Number of Years of Data = 5	
Parameter Assumptions	OLS Estimator	EB Estimator	OLS Estimator	EB Estimator	OLS Estimator	EB Estimator
Baseline assumptions						
$\rho_\varepsilon = 0.92$; $\rho_\omega = 0.03$; Single Subject Test	.36/.20	.41/.30	.26/.10	.30/.14	.20/.06	.24/.08
Sensitivity analysis for ICC parameters:						
$\rho_\varepsilon = 0.80$; $\rho_\omega = 0.03$	.35/.12	.37/.15	.25/.05	.27/.06	.19/.03	.20/.03
$\rho_\varepsilon = 0.92$; $\rho_\omega = 0$	.31/.12	.35/.17	.20/.05	.22/.06	.14/.02	.16/.03
$\rho_\varepsilon = 0.87$; $\rho_\omega = 0.08$	.39/.25	.44/.36	.31/.14	.36/.21	.26/.10	.30/.14
Sensitivity analysis for multiple subject tests						
$d = 2$ Tests; $\rho_d = 0.30$	.34/.18	.39/.26	.24/.08	.27/.11	.18/.04	.21/.06
Sensitivity analysis for longitudinal gain score data (two grades)						
$\rho_{t,t-1} = -0.38$	.35/.19	.40/.28	.25/.09	.29/.13	.19/.05	.23/.07
Sensitivity analysis for inclusion of pretest scores as a covariate						
$R^2_\omega = R^2_\varepsilon = 0.17$	.34/.18	.40/.27	.24/.08	.28/.12	.18/.05	.21/.06

Note. ICCs = Intraclass correlations. In cells with two entries, the first entry represents the Type I and II error rates that are restricted to be equal, and the second entry represents the overall false positive and false negative error rates that are restricted to be equal. For the sensitivity analysis for the ICC parameters, if changes in the values of the two indicated parameters do not offset each other, then the changes are assumed to be offset by changes in ρ_θ. The figures correspond to a scheme where a teacher is compared to the district average with 50 teachers in the district.

[a] See Figure 1 in the text for a depiction of this threshold value, which is measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

TABLE 6

School-Level Analysis: System Error Rates That are Restricted to Be Equal, by Threshold Value

	Threshold Value (Gain Score *SD*s From the Average)[a]					
	Type I = Type II Error Rate			Overall False Positive = Overall False Negative Error Rate		
Number of years of available data per school	.05	.1	.15	.05	.1	.15
Compare a school to the district average (5 schools in the district)						
1	.37	.26	.16	.14	.10	.06
3	.29	.13	.05	.07	.03	.01
5	.23	.07	.01	.04	.01	.00
10	.15	.02	.00	.02	.00	.00
Compare a school to the district average (30 schools in the district)						
1	.38	.28	.19	.16	.11	.07
3	.30	.15	.06	.08	.04	.01
5	.25	.09	.02	.05	.02	.00
10	.17	.03	.00	.02	.00	.00

Note. See the text for formulas and assumptions. Calculations assume test score data from a single subject area. Figures are based on the OLS estimator. See the text for formulas and assumptions.

[a] See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

of 1 to 3 percentage points. Thus, our benchmark findings are very close to the error rates implied by a quasi-gain model.

Finding 4: Using the OLS estimator and comparable threshold values, the school-level analysis will yield error rates that are about 5 to 10 percentage points smaller than for the teacher-level analysis. With 3 years of data, the OLS estimator for comparing a school to the district average would yield a Type I or II error rate of about 15% using a threshold value of .1 *SD*s—which is equivalent to setting the threshold for defining a high-performing school at the district's 83rd percentile school (Table 6). The corresponding error rate for FPR_TOT or FNR_TOT is about 4% (Table 6). Under this scenario, about 4 years of data would be required to achieve conventional Type I and II error rates of $\alpha = .05$ and $1 - \beta = 0.20$, and only 1 year would be required to achieve values of FPR_TOT = 0.05 and FNR_TOT = 0.20 (Table 7).

The school-level OLS analysis has more statistical power than the teacher-level OLS analysis, because school-level gain scores are estimated more precisely due to larger classroom and student sample sizes. Crucially, these

TABLE 7
School-Level Analysis: The Number of Years of Data Required to Achieve Various System Error Rates

Type I Error Rate/ Overall False Positive Rate	Type II Error Rate/Overall False Negative Rate			
	.05	.10	.15	.20
Threshold Value = .05 *SD*s[a]				
.05	31/5	24/3	21/3	18/2
.10	24/3	19/2	15/2	13/1
.15	21/2	15/2	12/1	10/1
.20	18/2	13/1	10/1	8/1
Threshold Value = .1 *SD*s[a]				
.05	8/2	6/2	5/2	4/1
.10	6/2	5/1	4/1	3/1
.15	5/1	4/1	3/1	3/1
.20	4/1	3/1	3/1	2/1
Threshold Value = .15 *SD*s[a]				
.05	3/2	3/1	2/1	2/1
.10	3/1	2/1	2/1	1/1
.15	2/1	2/1	1/1	1/1
.20	2/1	1/1	1/1	1/1

Note. In cells with two entries, the first entry represents the number of years required to achieve Type I and Type II error rates represented, respectively, by the row and column headers; the second entry represents the number of years required to achieve overall false positive and false negative rates represented, respectively, by the row and column headers. Figures are based on the ordinary least squares (OLS) estimator and assume test score data from a single subject area. The figures correspond to a scheme where a school is compared to the district average with 30 schools in the district. See the text for formulas and assumptions.

[a] See Figure 1 in the text for a depiction of these threshold values, which are measured in *SD*s of gain scores below or above the average true value-added measure in the appropriate population.

precision gains occur because the variances of the OLS estimator are conditional on fixed values of θ_{jk} and ψ_k. Statistical precision for the school-level analysis is much lower for the EB estimator due to the variance contribution of θ_{jk} (not shown).

Summary and Conclusions

This article has addressed likely error rates for measuring teacher and school performance in the upper elementary grades using student test score gain data and value-added models. This is a critical policy issue due to the increased interest in using value-added estimates to identify high- and low-performing instructional staff for special treatment, such as rewards and sanctions. Using rigorous statistical methods and realistic performance measurement schemes, the article presents evidence that value-added estimates for teacher-level analyses are subject to a considerable degree of random error when based on the amount of data that are typically used in practice for estimation.

Type I and II error rates for teacher-level analyses will be about 26% if 3 years of data are used for estimation. This means that in a typical performance measurement system, one in four teachers who are truly average in performance will be erroneously identified for special treatment, and one in four teachers who differ from average performance by 3 months of student learning in math or 4 months in reading will be overlooked.

Type I and II error rates pertain to specific teachers who have the greatest probability of being misclassified—those whose true performance is at the boundary between different performance ranges that merit different types of policy action. When the focus is on *overall* false positive and negative error rates for the full population of teachers who can be included in the calculations, rates of misclassification are lower. For example, with 3 years of data, overall misclassification rates will be about 10%.

Our results are largely driven by findings from the literature and new analyses that more than 90% of the variation in student gain scores is due to the variation in *student-level* factors that are not under the control of the teacher. Thus, multiple years of performance data are required to reliably detect a teacher's true long-run performance signal from the student-level noise. In addition, our analyses likely *understate* the error rates that would arise in practice because the analyses ignore nonrandom sources of error, such as nonrandom sorting of students to classrooms and schools or misspecification of the estimation model.

Our results strongly support the notion that policymakers must carefully consider system error rates in designing and implementing teacher performance measurement systems that are based on value-added models. In particular, policymakers should use judgment in identifying a range of tolerable error rates and then require each teacher's evaluation to be based on sufficient amounts of data for system error rates to lie within the tolerable range. With the hypothesis testing approach considered in this article, teacher evaluation systems that rely solely on value-added measures should use more than 3 years of data per teacher to achieve overall misclassification rates below 10%. If value-added measures are used in high-stakes personnel decisions—such as tenure—that need to be made on a shorter time-frame than that required for tolerable error rates, it is especially important to undertake implementation strategies, discussed below, that can further mitigate these error rates.

A performance measurement system at the *school* level will likely yield error rates that are about 5 to 10 percentage points lower than at the teacher level. This is because school-level mean gain scores can be estimated more precisely due to larger student sample sizes. Thus, current policy proposals to use value-added models for school-level accountability ratings may hold promise from the perspective of statistical precision. An important caveat, however, is that estimates of performance differences between schools could be biased, due, for instance, to nonrandom student sorting across schools.

Our findings highlight the need to mitigate system error rates. Misclassification rates could be lower if value-added measures were carefully coordinated with other measures of teacher quality. For instance, value-added estimates may serve as an initial performance diagnostic that identifies a *potential* pool of teachers warranting special treatment. While our findings suggest that some teachers would be erroneously identified during this initial round, a subsequent round of more intensive performance measurement focused on this pool could further separate those who do and do not deserve special treatment. Indeed, Jacob and Lefgren (2008) find that value-added measures and principals' assessments of teachers, in combination, are more strongly predictive of subsequent teacher effectiveness than each type of measure alone.

System error rates may be reduced further through a number of implementation strategies. For instance, developing tests with higher reliability, balancing student characteristics across classrooms, and assigning each teacher to multiple classes per year could improve the precision of teacher value-added estimates. With these and other strategies, value-added measures could be a less error-prone component of an overall "toolbox" for performance measurement.

It is important to recognize that our findings pertain to a prevalent class of estimators and performance measurement schemes that conduct hypothesis tests based on value-added estimates used in isolation. However, spurred by the recent infusions of funding from the federal government and private foundations, the development and application of teacher performance measures are ongoing and evolving, as districts have begun to explore new ways of combining value-added measures with other types of performance measures. Further research is warranted to determine the error rates generated by these and other schemes.

Although this article has focused on misclassification errors for teachers and schools, policymakers intend for educator evaluations ultimately to improve student outcomes. If personnel decisions are based on performance measures, then the accuracy of performance classifications will determine the mix of teachers who are retained, tenured, promoted, and fired, which in turn will affect the distribution of student outcomes (see, e.g., Bill & Melinda Gates Foundation, 2010). A direction for future research is to model and quantify the ways in which performance measurement systems with varying levels of error rates can lead to short- and long-run changes in the distribution of student outcomes.

Notes

1. Some evidence indicates that the quasi-gain model leads to less bias than the gain-score model when applied to nonexperimental data (Andrabi, Das, Khwaja, & Zajonc, 2011). However, to focus on precision, our article abstracts away from nonexperimental sources of bias.

2. We use the district mean as the cutoff value of the null hypothesis for several reasons. First, it is the cutoff value that is used in practice and is transparent. Second, defining a cutoff value that, instead, is located at the performance level of a very high-performing (or low-performing) teacher could lead to unacceptably high false negative error rates, as defined later in this article. Although it might be possible to define an "optimal" cutoff value that minimizes a social loss function, we do not consider this approach because it requires a subjective loss function specification and has not yet been implemented in practice.
3. We express *SD*s in gain score units because the HLM pertains to gain scores, but the results would be identical if the *SD*s (and our *SD* targets) were instead expressed in posttest score *SD* units.
4. The studies are Goldhaber and Hansen (2008); Hanushek, Kain, O'Brien, and Rivkin (2005); Kane, Rockoff, and Staiger (2008); Kane and Staiger (2008); Koedel and Betts (2009); McCaffrey, Sass, Lockwood, and Mihaly (2009); Nye, Konstantopoulos, and Hedges (2004); Rivkin, Hanushek, and Kain (2005); Rockoff (2004); and Rothstein (2010). The data for the primary analysis come from national evaluations of teachers from Teach for America, teachers from alternative certification programs, early elementary math curricula, reading and mathematics software products, and reading comprehension interventions.

References

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

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics*, *3*, 29–54.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate. A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 1289–1300.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. New York, NY: Springer-Verlag.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf. Seattle, WA.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions*. New York, NY: MDRC.

Bloom, H., Richburg-Hayes, L., & Black, A. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30–59.

Gelman, A., Hill, J., & Yajima, M. (2009). *Why we (usually) don't have to worry about multiple comparisons. Department of statistics working paper*. New York, NY: Columbia University.

Goldhaber, D., & Hansen, M. (2008). Is it just a bad class? Assessing the stability of measured teacher performance. *CRPE working paper 2008_5.* Seattle, WA: Center on Reinventing Public Education.

Hanushek, E., Kain, J., O'Brien, D., & Rivkin, S. (2005). The market for teacher quality. *NBER working paper 11154*. Cambridge, MA: National Bureau of Economic Research.

Hanushek, E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings*, *100*, 267–271.

Harris, D., & Sass, T. (2006). Value-added models and the measurement of teacher quality. *Working paper*. Tallahassee: Florida State University.

Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

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

Kane, T. (2004). The impact of after-school programs: Interpreting the results of four recent evaluations. *Working paper*. Los Angeles, CA: University of California.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*, 615–631.

Kane, T., & Staiger, D. (2002a). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*, 91–114.

Kane, T., & Staiger, D. (2002b). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy: 2002*. Washington, DC: Brookings Institution Press.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER working paper* 14607. Cambridge, MA: National Bureau of Economic Research.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley.

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

Koedel, C., & Betts, J. (2009). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Working paper*. Columbia: University of Missouri.

Lindley, D.V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistics Society, Series B*, *34*, 1–41.

Lipscomb, S., Teh, B., Gill, B., Chiang, H., & Owens, A. (2010). *Teacher and principal value-added: Research findings and implementation practices*. Cambridge, MA: Mathematica Policy Research.

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

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

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

National Institute for Excellence in Teaching. (2009). *TAP: The system for teacher and student advancement*. Retrieved October13, 2009, from http://www.tapsystem.org/

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage.

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

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review (AEA Papers and Proceedings)*, *94*, 247–252.

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

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

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

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

Solmon, L., White, J. T., Cohen, D., & Woo, D. (2007). *The effectiveness of the teacher advancement program*. Santa Monica, CA: National Institute for Excellence in Teaching.

Springer, M., Ballou, D., & Peng, A. (2008). Impact of the teacher advancement program on student test score gains: Findings from an independent appraisal. *Working paper* 2008-19. Nashville, TN: National Center on Performance Incentives.

Todd, P., & Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, F3–F33.

U.S. Department of Education. (2010). *Race to the top application for initial funding: Tennessee*. Retrieved May 21, 2010, from http://www2.ed.gov/programs/racetothetop/phase1-applications/tennessee.pdf

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

Authors

PETER Z. SCHOCHET, Ph.D., is a Senior Fellow at Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393; email: pschochet@mathematica-mpr.com. His research interests include causal inference underlying experiments, statistical power, clustered designs, regression discontinuity designs, propensity score matching designs, multiple testing, mediator analyses, value-added modeling, and conducting impact evaluations of education, employment, and welfare programs.

HANLEY S. CHIANG, Ph.D., is a Researcher at Mathematica Policy Research, 955 Massachusetts Ave., Suite 801, Cambridge, MA 02139; email: hchiang@mathematica-mpr.com. His research interests include program evaluation, teacher quality, and incentives in public education.

Manuscript received May 2, 2011
Revision received August 25, 2011
Accepted October 21, 2011