Measuring the Power of Learning.®

Research Report
ETS RR–17-25

# Measurement Error and Bias in Value-Added Models

Michael T. Kane

December 2017

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Measurement Error and Bias in Value-Added Models

Michael T. Kane

Educational Testing Service, Princeton, NJ

By aggregating residual gain scores (the differences between each student's current score and a predicted score based on prior performance) for a school or a teacher, value-added models (VAMs) can be used to generate estimates of school or teacher effects. It is known that random errors in the prior scores will introduce bias into predictions of the current scores, and thereby, into the estimated residual gain scores and VAM scores. The analyses in this paper examine the origins of this bias and its potential impact and indicate that the bias is an increasing linear function of the student's prior achievement and can be quite large (e.g., half a true-score standard deviation) for very low-scoring and high-scoring students. To the extent that students with relatively low or high prior scores are clustered in particular classes and schools, the student-level bias will tend to generate bias in VAM estimates of teacher and school effects. Adjusting for this bias is possible, but it requires estimates of generalizability (or reliability) coefficients that are more accurate and precise than those that are generally available for standardized achievement tests.

**Keywords** VAMs; bias; residual gain scores; bias in school-level VAM scores; validity; generalizability of state test scores

doi:10.1002/ets2.12153

Teachers and schools have always been evaluated in terms of how well they seemed to be functioning. Traditionally, the evaluations have tended to focus on input variables like what is being taught, the methods being used in teaching, the perceived quality of materials and teacher performance, and the overall climate of the school or class. Starting with the No Child Left Behind legislation and the development of value-added models (VAMs), the focus has shifted to the evaluation of outcome measures, particularly student performance on standardized tests (Koretz & Hamilton, 2006).

VAMs are designed to support inferences about a school's or teacher's contribution to student growth over the course of a year by comparing the current test scores of their students to the scores of the same students in the prior year (or years), either directly (in gain-score models) or indirectly (e.g., in residual gain score or covariate-adjustment models). The intent is to isolate the gains in student achievement (as indicated by changes in test scores from one year to the next) that are attributable to the school or teacher for the given year and to use some function of these gains in the evaluation of the teacher or school. VAMs employ sophisticated statistical models to control for various factors that could have an impact on student scores, and by controlling for these extraneous factors, they aim to get a good estimate of teacher or school effects (Braun, Chudowsky, & Koenig, 2010; McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

## Residual Gain Scores

A natural way to evaluate student growth along some dimension would be to estimate the student's standing on the dimension at the current time and at some prior time and to examine the change, or growth, between the current and prior estimates. Growth scores are familiar, simple, and direct, but the current and prior scores have to be on the same scale (or on vertically aligned scales) for the differences to make sense. Furthermore, the scale has to be an interval scale in the sense that a difference of a certain number of points has, at least approximately, the same meaning along the scale, so that it makes sense to compare gain scores from different parts of the scale (Braun et al., 2010; Briggs, 2013; Haertel, 2013; Kolen, 2006; Koretz & Hamilton, 2006; Rothstein, 2009). Unfortunately, it is very hard to put the test scores from different grade levels on a common scale and to demonstrate that the resulting scale is an interval scale (Briggs, 2013); some uncertainty about scale characteristics is not a problem for many applications of vertical scaling, but it is a serious problem if the proposed use of the scores (e.g., educational accountability based on growth scores) demands that the vertical scale be demonstrably equal interval (Ballou, 2009; Betebenner, 2009; Briggs & Domingue, 2013; McCaffrey et al., 2003).

*Corresponding author:* M. Kane, E-mail: mkane@ets.org

One way to remove the need to establish an interval scale across grades is to use residual gain scores (Castellano & Ho, 2013) or covariate-adjustment models (McCaffrey et al., 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). In these models, student test scores for the current year are adjusted to remove the effects of prior learning, as indicated by the prior year's scores and other available student and context variables. In the simplest version, a student's prior test score is used to predict the current test score, and the difference between the actual and predicted scores, the *residual*, is interpreted as an indicator of the student's growth relative to the expected growth of students with the same predicted score. By aggregating the results across the students in a school or taught by a teacher, an inference can be drawn about the effectiveness of the school or teacher. The accuracy of these inferences will depend on the completeness and accuracy of the prior information and on the accuracy of the statistical model in making adjustments for extraneous factors (recognized and unrecognized).

The logic of ordinary least squares (OLS) regression does not require that the scores on the two tests be on the same scale. The predicted current scores are on the current-score scale, because regression-based predictions are on the scale of the dependent variable, regardless of the independent variables included; for example, OLS regression can be used to predict students' heights in inches based on their ages in months. So, in using residual gain scores, it is not necessary to link last year's scores to this year's scores in a vertical scale. However, it is still necessary to assume that a difference of a certain number of points has more or less the same meaning along the score scale for the current test scores (i.e., that the scaled scores for each grade are approximately interval, but the grade scales are not necessarily linked).

By focusing on the residual score that results when the predicted current score is subtracted from the actual current score, prior educational experiences (as reflected in the prior year's scores) are controlled to some degree, but such adjustments can be biased to the extent that the predicted score does not include all factors that may have an impact on student performance. Bias can also result from errors of measurement in the prior scores included in the prediction equation. Estimates of residual gain scores rely on OLS regression, and OLS regression assumes that the independent variables are measured without error. It is well known that random errors in an independent variable will lead to bias in the estimation of the true-score relationship between the dependent variable and the independent variable (Berry, 1993; Campbell & Kenny, 1999; Harris, 1963; Lockwood & McCaffrey, 2014; McCaffrey et al., 2003). The analyses in this paper examine how this bias gets incorporated into student gain scores and then into teacher and school effects, and the results indicate that the bias due to random error in the prior scores can be substantial.

As indicated later in this paper, the resulting bias is an increasing linear function of the student's prior true scores; the gains for students with high true scores on the prior year's test will be overestimated, and the gains for students with low true scores in the prior year will be underestimated. To the extent that students with relatively low and high true scores tend to be clustered in particular classes and schools, the student-level bias will generate bias in estimates of teacher and school effects. If not corrected, this source of bias could have a substantial negative impact on estimated VAM scores for teachers and schools that serve students with low prior true scores and could have a substantial positive impact for teachers and schools that serve mainly high-performing students. Given that the results of VAMs may be used for high-stakes decisions about teachers and schools in the context of accountability programs (Braun et al., 2010; Winters & Cowen, 2013), any substantial source of bias would be a matter of great concern.

The discussion, presented below, of bias in residual gain scores based on OLS regression does not necessarily apply to VAMs that do not rely on such residual gain-scores. Analyses based on simple growth scores, as such, would not be subject to this bias. Student growth percentiles (Betebenner, 2009) are based on quantile regression, which is different from OLS regression in several ways (e.g., focusing on current-score percentiles rather than means and residuals), but like OLS regression, quantile regression also assumes that the independent variable (the prior year scores) is free of measurement error. As a result, student growth percentiles are likely to include bias due to random errors in the prior scores, but this possibility is not addressed here (see McCaffrey, Castellano, & Lockwood, 2015).

## Bias in Residual Gain Scores

The use of OLS regression introduces bias into the estimation of residual gain scores if the prior scores contain random error. To illustrate the direction and potential magnitude of this source of bias, I will examine a particularly simple special case. I will assume that the current true scores are a linear function of the previous year's true scores for all students:

$$T_Y = \alpha + \beta T_X, \tag{1}$$

where $T_Y$ is a student's current true score, and $T_X$ is the student's true score from the prior year. The true scores are expected values of the $X$ and $Y$ scores for each student over independent replications of the testing procedure. For the kinds of achievement tests used in VAMs, replications could involve different items, administered on different days during some part of the school year, and for constructed responses, scored by different raters.

The values of the constants, $\alpha$ and $\beta$, will depend on the scaling employed in the testing program. If some kind of vertical scaling is employed in the testing program, with score means increasing from year to year, $\alpha$ and $\beta$ would both generally be positive, and their values would depend mainly on the details of the scaling procedure and on the generalizability/reliability of the observed scores in each year. For the purposes of this paper, the value of $\alpha$ is largely irrelevant, and the precise value of $\beta$ is not critical, as long as it is not zero. It is convenient and highly plausible to assume that $\beta$ is positive; test scores are highly correlated across years (Boyd, Langford, Loeb, & Wyckoff, 2013).

In state testing programs, the scores at each grade level are commonly scaled to have a common mean and common variance. If the observed scores are scaled to have the same mean and variances for each grade level and their reliabilities are approximately the same across grade levels, $\alpha$ would be very close to 0.0, and $\beta$ would be very close to 1.0.

Equation 1 implies that teachers and schools have no differential impact on student growth. Students' current true scores have the same relationship to the student's previous true scores, independent of what school students attend or what teachers they have. That is, we are assuming that all teachers and schools are uniformly good or bad, and therefore, any systematic differences between estimated current scores and actual current scores for teachers or schools can be interpreted as bias in estimates of school or teacher effects.

Equation 1 does not provide a realistic model for changes in true scores from 1 year to the next, and this hypothetical relationship is introduced only to facilitate an examination of the bias in residual gain scores due to random errors in the prior year's scores. A particularly simple model for which we know what the outcomes should be has been postulated as a *thought experiment*. If the residual gain estimates yield the expected results, our confidence in the use of residual-gain scores to estimate differential learning across students and average student performance across teachers would increase. However, if the residual gain estimates do not yield the expected results (i.e., no differential change), we can conclude that these analyses are biased, at least in some cases.

The observed scores from the previous year have measurement errors, and can be represented as:

$$X = T_X + \varepsilon_X, \tag{2}$$

where $T_X$ is the student's true prior score, and $\varepsilon_X$ is the random error in the student's prior score; the error has a mean of zero, and is uncorrelated with the true score for $X$. Similarly, the observed scores from the current year also have error, and can be represented as:

$$Y = T_Y + \varepsilon_Y, \tag{3}$$

where $T_Y$ is the student's true current score, and $\varepsilon_Y$ is the current-score random error, which has a mean of zero and is uncorrelated with the true score for $Y$. The random errors, $\varepsilon_X$ and $\varepsilon_Y$, are also assumed to be uncorrelated with each other and with the two true scores.

A student's scores on an achievement test would be likely to fluctuate if the testing were replicated with different samples of tasks, on different occasions (within some testing window), and for extended-response tasks, for different scorers, and these fluctuations, which are uncorrelated with all other variables (not defined in terms of the observed scores), are generally treated as random errors, or noise (Brennan, 2001a; Kane, 2011). Taking the true scores to be expected values over replications, as is done in classical test theory and generalizability theory, the errors have a mean of 0.0 and are uncorrelated with true scores. In classical test theory, the error is generally treated as a single random variable; in generalizability theory, the error is analyzed into separate components representing different sources of variability that contribute to the overall error (Brennan, 2001a, 2001b). The error terms in Equations 2 and 3 are assumed to include the contributions of all sources of random error (fluctuations with a mean of zero and zero correlations with the other variables); indices that involve only one or two of the relevant sources of error (e.g., indices like coefficient alpha) provide underestimates of this overall error.

In this section, we examine the impact of the overall error in the prior observed scores on estimates of residual gain scores. In a later section, we consider the magnitudes of different sources of error for state achievement tests, and their contribution to the overall error.

The expected value of the observed $X$ scores over all students in the population is:

$$\mu_X = E(X) = E\left(T_X + \varepsilon_X\right) = E\left(T_X\right). \tag{4}$$

The variance in the observed $X$ scores is:

$$\sigma_X^2 = E\left(X - \mu_X\right)^2 = E\left(T_X + \varepsilon_X - \mu_X\right)^2 = \sigma_{T_X}^2 + \sigma_{\varepsilon_X}^2. \tag{5}$$

The generalizability/reliability of the prior test scores, $\rho_{XX'}$, is the ratio of true-score variance to observed-score variance.

Similarly, the expected value of the observed $Y$ scores over the population is:

$$\mu_Y = E(Y) = E\left(T_Y + \varepsilon_Y\right) = E\left(T_Y\right). \tag{6}$$

Using Equations 1 and 4, it is also equal to:

$$\mu_Y = E\left(T_Y + \varepsilon_Y\right) = E\left(\alpha + \beta T_X + \varepsilon_Y\right) = \alpha + \beta\mu_X. \tag{7}$$

The variance in $Y$ is given by:

$$\sigma_Y^2 = E\left[T_Y - \mu_Y + \varepsilon_Y\right]^2 = \sigma_{T_Y}^2 + \sigma_{\varepsilon_Y}^2, \tag{8}$$

and using Equations 1 and 7, it is also equal to:

$$\sigma_Y^2 = E\left[\beta\left(T_X - \mu_X\right) + \varepsilon_Y\right]^2 = \beta^2\sigma_{T_X}^2 + \sigma_{\varepsilon_Y}^2. \tag{9}$$

The covariance between $Y$ and $X$ is given by:

$$\begin{aligned}\operatorname{cov}(Y, X) &= E\left[\left(Y - \mu_Y\right)\left(X - \mu_X\right)\right] = E\left[\left(\beta T_X - \beta\mu_X + \varepsilon_Y\right)\left(T_X + \varepsilon_X - \mu_X\right)\right] \\ &= \beta\sigma_{T_X}^2.\end{aligned} \tag{10}$$

The correlation between $Y$ and $X$ is:

$$\rho_{XY} = \frac{\operatorname{cov}(Y, X)}{\sigma_Y\sigma_X} = \frac{\beta\sigma_{T_X}^2}{\sigma_Y\sigma_X}. \tag{11}$$

Consider a student with a true score of $T_X$ in the prior grade. By our assumption that students' current true scores are linearly related to their prior true scores, this student would have an observed score in the current grade of:

$$Y = T_Y + \varepsilon_Y = \alpha + \beta T_X + \varepsilon_Y. \tag{12}$$

Using a general form of an OLS regression, employing population values for the parameters, the student would have a predicted score for the current grade of:

$$\widehat{Y} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}\left(X - \mu_X\right) + \mu_Y. \tag{13}$$

In value-added contexts, the numbers of students used to estimate population statistics (means, standard deviations, correlations) are typically quite large, and therefore, the estimates should be very close to the population values, and I have not bothered to complicate the notation to indicate that the current-score estimates would be based on an estimated regression equation.

In addition, we have assumed that the only variable used to predict students' current scores is their prior scores. In some cases, more than one prior score may be used, but these prior scores can generally be combined into a single prior score with suitable weighting; the combined score would generally have a higher generalizability than a single prior score would have, and the magnitude of the problem discussed here would be reduced, but it would not be eliminated. Also, additional student variables may be included in the prediction, but in these cases, the prior test scores tend to dominate the prediction equation (Haertel, 2013; McCaffrey et al., 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rothstein, 2009; Sanders & Horn, 1998), and the simple model employed here provides an indication of how the bias due to random errors in prior scores would operate in more complex models with student background variables as additional covariates.

The slope of the regression line in Equation 13 can be simplified a bit using Equation 11:

$$\rho_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{\beta\sigma^2_{T_X}}{\sigma_Y\sigma_X}\frac{\sigma_Y}{\sigma_X} = \beta\frac{\sigma^2_{T_X}}{\sigma^2_X} = \beta\rho_{XX'}, \tag{14}$$

where $\rho_{XX'}$ is the generalizability of the prior test scores, the ratio of prior true-score variance to the prior observed-score variance. This generalizability coefficient is intended to reflect all sources of random fluctuations contributing to the overall error.

The predicted current score as a function of the prior score can then be written as:

$$\hat{Y} = \beta\rho_{XX'}\left(X - \mu_X\right) + \mu_Y. \tag{15}$$

Note that the slope of this OLS regression equation is different from the slope of the true-score relationship, given by:

$$T_Y = \beta\left(T_X - \mu_X\right) + \mu_Y = \alpha + \beta T_X. \tag{16}$$

The slope of the estimated OLS regression line in Equation 15 is smaller than the slope of the line representing the true-score relationship in Equation 16 by a factor of the generalizability of the prior scores. This effect is an example of regression to the mean (Campbell & Kenny, 1999).

Using Equations 2, 3, 7, 15, and 16, the residual gain score can be written as:

$$Y - \hat{Y} = \left[\alpha + \beta T_X + \varepsilon_Y\right] - \left[\beta\rho_{XX'}\left(T_X + \varepsilon_X - \mu_X\right) + \left(\alpha + \beta\mu_X\right)\right]. \tag{17}$$

Assuming, as we are, that there are no teacher or school effects, and the current scores are adjusted for prior learning (specifically, the prior test scores), the expected value of the residuals in Equation 17 (over independent replications of testing procedures) should be zero for all students.

The expected value, over replications of the testing procedure, of the residual, given $T_X$, is the expected value, over replications, of $(Y - \hat{Y})$, given $T_X$:

$$E\left(Y - \hat{Y}|T_X\right) = \left[\alpha + \beta T_X\right] - \left[\beta\rho_{XX'}\left(T_X - \mu_X\right) + \left(\alpha + \beta\mu_X\right)\right], \tag{18}$$

or

$$E\left(Y - \hat{Y}|T_X\right) = \beta\left(1 - \rho_{XX'}\right)\left(T_X - \mu_X\right). \tag{19}$$

Note that the expected residual in Equation 19 is not generally equal to zero, unless the reliability of the prior scores is 1.0, or the particular prior true score under consideration, $T_X$, equals the mean of the prior true scores. Assuming that $\beta$ is positive (i.e., that the observed scores and therefore the true scores are positively correlated from year to year), the expected residual is positive for prior true scores above the mean and gets larger as the prior true score gets higher, and the expected residual is negative for prior true scores below the mean and gets larger in magnitude as the prior true score gets lower.

The direction of the bias is consistent with empirical findings indicating that, even after controlling for prior scores and some student characteristics, teachers of low-scoring students are at a disadvantage:

> … we found that students' residualized achievement scores were, in most analyses, more strongly predicted by the students' prior achievement and the course they were in than by the teacher …. Each teacher appeared to be significantly more effective when teaching upper-track courses than the same teacher appeared when teaching lower-track courses. (Newton et al., 2010, p. 18)

The bias associated with measurement error in the prior scores can explain these empirical results, at least in part.

## Magnitude of the Bias in Individual-Student Residual Gain Scores

As noted above, the bias associated with random errors in the prior year's test scores is not a constant; it is equal to zero for students with prior scores equal to the prior mean score and is increasingly positive for higher prior true scores and increasingly negative for lower prior true scores. The bias is also zero if the prior test scores have a generalizability coefficient of 1.0, and otherwise, its magnitude gets larger as the generalizability of the prior scores decreases.

Bias tends to be a more serious problem than random errors, but it is not necessarily so serious if the magnitude of the bias is small compared to other sources of random and systematic error. So it is useful to get some indication of the magnitude of the bias associated with random errors in the prior scores. The bias due to errors in the prior scores will potentially apply at the individual student level and at both the teacher and school levels. In this section, I will focus on the individual student bias. In the next section, I will focus on the school level; the effect is more complicated at the teacher level, and empirical estimates of some of the relevant quantities needed to estimate the effect are not readily available at the teacher level.

As noted earlier, in state testing programs in which scale scores at each grade level have fixed means and have similar standard deviations from year to year, $\beta$ is likely to be quite close to 1, and the bias introduced by the errors of measurement in the prior scores, given by Equation 19, can be simplified by assuming that $\beta = 1$:

$$E\left(Y - \widehat{Y}|T_X\right) = \left(1 - \rho_{XX'}\right)\left(T_X - \mu_X\right). \tag{20}$$

The first factor in Equation 20 reflects the proportion of the observed variance in the prior scores that is attributable to random error (as indicated by the generalizability coefficient), but the generalizability coefficient under consideration here is not simply an indicator of variability over different samples of items (as coefficient alpha would be). Rather, the random errors that are of concern include all random sources of variability in the prior test scores, which are uncorrelated with the prior true scores and the current true scores. This generalizability coefficient could be estimated by a generalizability coefficient (Brennan, 2001b; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) that reflects all potential sources of error (e.g., variability associated with the sampling of items and occasions, and if relevant, any variability over raters).

As indicated earlier, the estimated generalizability/reliability coefficients for state tests (generally coefficient alpha) are likely to overestimate the actual generalizability reflecting the overall random error. Any aspects of testing contributing to score variability that is uncorrelated with the true scores and with other variables of interest are source of random error. In generalizability theory (Brennan, 2001b; Cronbach et al., 1972), each identifiable source of variance (e.g., sampling of items) is referred to as a *facet*, and we can have item facets, occasion facets, rater facets, and so on. The contribution of each facet to the overall error depends on the magnitudes of variance components associated with the facet, the data-collection design, and the proposed interpretation of the scores. Because commonly used estimates of the reliability of standardized test scores (e.g., coefficient alpha) tend to focus on only one source of error, they tend to underestimate the total error and to overestimate generalizability coefficients that incorporate all potentially substantial sources of error. The variability associated with occasions (Cronbach, Linn, Brennan, & Haertel, 1997; Lane & Stone, 2006) will reflect changes in student motivation, wellness, and so on, as well as any variability in the environment and conditions of test administration from one occasion to another, and it is potentially substantial.

The reported estimates of generalizability (generally reported as reliability coefficients) for state tests are typically based on analyses of internal homogeneity (e.g., coefficient alpha) and tend to be about 0.90 or a bit higher. For example, the North Carolina Department of Public Instruction (2014) reported coefficient alphas between 0.88 and 0.92 in English language arts (ELA) and between 0.91 and 0.93 in mathematics for their end-of-grade assessments for grades 3 – 8. The New York State Education Department (2015) reported alphas for grades 3 – 8 ranging from 0.89 to 0.92 for ELA and, ranging from 0.93 to 0.95 for mathematics. The Wisconsin Department of Public Instruction (2015) reported coefficient alphas for ELA ranging from 0.87 to 0.91, and for mathematics, ranging from 0.89 to 0.91. Ferrara (2006) reviewed technical reports from 11 states (not including the three mentioned above) to evaluate the technical characteristics of state testing programs and reported that all provided internal-consistency estimates (coefficient alpha) for content-area (e.g,, reading, mathematics) scores (with one reporting a stratified alpha); most of the alphas were greater than 0.85 and ranged into the low to mid 0.90s.

Given that these values are likely to be overestimates, the generalizability coefficient in Equation 21 is likely to be less than 0.90 for state testing programs and could easily be in the low 0.80s if the variance components for occasions, and (if appropriate) raters, in addition to tasks/items (the only source of error included in coefficient alpha), were included in the estimates. Estimates of generalizability coefficients that include occasions as a source of variance are rare for state tests. Rothstein (2009) referenced estimates of the test – retest reliability for the North Carolina seventh grade reading test reported by Sanford (1996) of 0.86. Boyd et al. (2013) suggested that the actual error in state achievement test scores may be twice as large as the reported values, which would suggest a generalizability/reliability coefficient of about 0.80. In this

section, I use 0.90 and 0.80 as benchmark values for the generalizability coefficient in examining possible magnitudes for the bias; the actual values for different state tests are probably distributed across this range.

Assuming a generalizability of 0.90, the first term in Equation 20 would be 0.10, and a student with a prior true score that is one true-score standard deviation above the prior mean true score would have a positive bias of a 10th of a true-score standard deviation in the prior scores, $X$, and a student with a prior true score one standard deviation below the mean would have a negative bias of a 10th of a true-score standard deviation in $X$. Note that, because of the assumption in Equation 1, these students would also have current true scores that are, respectively, one standard deviation above and below the current mean true score.

Similarly, students with prior true scores that are two standard deviations above or below the mean prior score would experience a positive or negative bias of 2/10ths of a true-score standard deviation in $X$, and students with true scores three standard deviations above or below the mean would experience a positive or negative bias of 3/10ths of a true-score standard deviation in $X$.

For a generalizability of 0.80, the bias would be twice as large as it would be for a generalizability of 0.90, and for students with prior true scores one true-score standard deviation above or below the mean prior score, the magnitude of the bias would be 2/10ths of a true-score standard deviation in $X$. For students with prior true scores two true-score standard deviations from the mean, the magnitude of the biases would be 4/10ths of a true-score standard deviation in $X$, and for students with prior true scores three standard deviations from the mean, the magnitude of the biases would be 6/10ths of a true-score standard deviation in $X$. So, the student-level bias can be quite substantial for students with prior true scores two or three true-score standard deviations from the mean prior score.

## Magnitude of the Bias in Residual Gain Scores for Schools

The bias in estimating residual gain scores, based on a fallible prior score, can be quite large (over half a true-score standard deviation) for individual students. If not corrected, this bias can generate bias in estimates of teacher and school effects, but the magnitudes of the teacher and school effects will tend to be considerably smaller than the biases in individual student estimates because of averaging. The positive biases for students with prior scores above the mean and the negative biases for students with prior scores below the mean will tend to cancel out. If the mean prior score for a teacher or school happens to be zero, there will be no bias attributable to errors of measurement in the prior scores. If the teacher or school mean prior score is different from zero, the errors of measurement in prior scores will have an impact on estimated teacher and school effects.

To get at least a rough indication of the impact of the bias to be expected at the school level, we need an estimate of the test-score generalizability and an indication of the standard deviation of the average true scores across schools. The expected value of the bias for a school will be a linear function of the average prior true score for the school:

$$E\left(Y-\widehat{Y}|\overline{T_X}\right)=\left(1-\rho_{XX'}\right)\left(\overline{T_X}-\mu_X\right). \tag{21}$$

As indicated above, the generalizability/reliability coefficients for state tests are likely to be between about 0.90 and about 0.80, and we can take the middle of this range, 0.85, as a reasonable estimate of the overall generalizability of the test scores.

We have less information on the variability of school means. The prior-score means can be different from zero for a number of reasons. If students were randomly assigned to schools, the distribution of prior-score means would have a mean equal to the prior score mean of the population and a standard deviation equal to the standard deviation of the individual prior scores divided by the square root of the number of students in the school. This random component will tend to be small for schools with large numbers of students, but could be significant for small schools. This random component would vary from year to year as new random samples of students are selected each year; over consecutive years, it functions as a source of random error in evaluating school performances, but for a given year, it functions as a source of bias in the sense, that if not corrected, it puts individual schools at an advantage or disadvantage before they begin the school year.

Students are not randomly assigned to schools (Aaronson, Barrow, & Sander, 2007; Briggs & Domingue, 2011; Stone & Lane, 2003). Assignment to schools is generally based, to a large extent on where students live, and neighborhoods tend to vary considerably in their income levels and other demographic variables, and test scores are correlated with these factors.

Stone and Lane (2003) reported that, between 1993 and 1997, the standard deviation across schools of mathematics scores on the Maryland State Performance Assessment Program (MSPAP) ranged between .45 and .48 of the standard deviation in individual scores, and the standard deviation across schools of reading scores ranged from .34 to .41 of the standard deviation in individual scores. If one takes 0.40 as a representative value for the standard deviation of observed scores and takes 0.85 to be the generalizability coefficient, the true-score standard deviation of the school means can be obtained by multiplying the observed-score standard deviation by the square root of the generalizability, yielding an estimate of about 0.37.

Based on data from North Carolina, Kane and Staiger (2002) concluded that:

> In North Carolina elementary schools near the national average in size (between sixty-five and seventy-five students with valid test scores), the variance in mean reading and math scores was 0.087 and 0.092 respectively. Dividing the estimated amount of variance due to sampling variation for a school of average size (0.013) by the total variance observed for such schools, we would infer that 14 to 15 percent of the variation in fourth-grade math and reading test scores was due to sampling variation. (p. 241)

The observed-score standard deviations is about 0.30 for both reading and mathematics, so, the true-score standard deviation for school means is about 0.27 for both areas, For smaller schools, with around 40 students, Kane and Staiger's (2002) analyses suggested that observed-score standard deviation across schools would be about 0.4, so the true-score standard deviation across schools would be about 0.37.

The two analyses are pretty similar, especially given that they involve tests and data sets from different states. Given these two analyses, it would seem reasonable to take the universe score variance to be about 0.32 (the average of 0.27 and 0.37). Substituting 0.85 for the generalizability in Equation 21,

$$E\left(Y-\widehat{Y}|\overline{T_X}\right)=.15\left(\overline{T_X}-\mu_X\right). \tag{22}$$

Given a school true-score standard deviation across schools of 0.32 student standard deviations and a roughly normal distribution, we can assume that about a third of the school means would be more than 0.32 student standard deviations above or below the overall mean and would therefore experience a bias due to error in prior scores of over $(0.15)(.32)=0.05$. Further, under these conditions, it would not be uncommon (about 5% of the schools) for the school means to be over 0.64 student standard deviations above or below the overall mean and to experience a bias due to error in prior scores of over 0.10.

The school level biases are much smaller than the student-level biases, but they are large enough to be a source of concern. As noted earlier bias tends to be more serious than random errors, because bias does not cancel out over time; rather it can accumulate over time. A bias of 0.05 or 0.10 is a serious concern in a context in which the true-score standard deviation is only about 0.30.

Teachers and schools serving students with low prior scores would be at a serious disadvantage, and teachers and schools serving students with high prior scores would be at an advantage. The bias discussed in this paper is a statistical artifact related to regression to the mean, but it is not unrelated to concerns about social bias defined in terms of race, gender, and socioeconomic level. Stone and Lane (2003) found that school means were strongly related to measures of socioeconomic status; in particular, they found that the percentage of students participating in free or reduced price lunch was consistently related to school MSPAP averages.

Boyd, Lankford, Loeb, Rockoff, and Wyckoff (2007) reported that for students in the New York City public schools in 2005, over 50% of fourth-grade students in the highest poverty decile (the poorest students) and over 75% of eighth-grade students in this decile failed to meet the proficiency standard on the end-of-year ELA exam, while only about 18% of fourth-grade students and about 41% of eight-grade students in the lowest poverty decile failed to meet this standard; high-poverty schools are likely to have students with low-prior scores.

## Correcting for the Bias due to Measurement Errors in Prior Scores

A number of methods have been proposed to correct for the bias due to random errors in the prior test scores by adjusting for the regression-to-the-mean produced by random errors in the prior scores (Fuller, 1987; Lockwood & McCaffrey, 2014). For example, if we adjust the slope of the OLS regression in Equation 15, by dividing it by the generalizability of the

prior scores, we get an unbiased estimate of the true-score relationship in Equation 1, and the bias due to random errors in the prior scores is eliminated. However, as noted earlier, most readily available estimates of the generalizability (e.g., coefficient alpha) do not include all known sources of error, and therefore, these estimates are likely to underestimate the overall error and overestimate generalizability coefficients.

If we correct for the error in the prior scores by dividing by an estimate of their generalizability, $r_{XX'}$, the expected value of the bias indicated by Equation 19 would be:

$$E\left(Y-\widehat{Y}|T_X\right)=\beta\left(1-\frac{\rho_{XX'}}{r_{XX'}}\right)\left(T_X-\mu_X\right). \tag{23}$$

If the estimated generalizability equals the generalizability of interest (i.e., if $r_{XX'}=\rho_{XX'}$), Equation 23 would be 0.0 for all values of the prior true score, and the bias due to random errors in the prior scores would disappear. If the estimated generalizability is greater than the generalizability of interest (i.e., if $r_{XX'}>\rho_{XX'}$), the bias would decrease in magnitude, but be in the same direction. If the estimated generalizability is smaller than the generalizability of interest (i.e., if $r_{XX'}<\rho_{XX'}$), the correction would result in a bias in the opposite direction, and could be larger or smaller than the original bias.

If one were to adequately correct for the bias due to random errors in the prior scores, one would need a reliability/generalizability coefficient that reflects all significant sources of random error in the prior scores. As indicated earlier, the reliability coefficients reported for state testing programs are likely to be overestimates of the generalizability coefficient of interest because they omit sources of error associated with variability over occasions, raters, and contexts of testing, and a more realistic estimate of the generalizability reflecting variability over tasks, occasions, raters, and contexts would probably be in the middle to low 0.80s (Boyd et al., 2013; Sanford, 1996). Assuming that the generalizability associated with the total error is 0.80, and the generalizability used to correct for the impact of random errors in the prior scores is 0.90, the bias would be reduced by about 50%. If the generalizability associated with the total error is 0.85, and the generalizability used to correct for the impact of random errors in the prior scores is 0.90, the bias would be reduced by about 63%. Correcting for the errors in prior scores, using a coefficient that underestimates the magnitudes of the errors (e.g., coefficient alpha) reduces the bias, but does not eliminate it.

Progress is being made on methods to minimize the bias introduced by random errors in the prior scores (Lockwood & McCaffrey, 2014), but most of these methods rely on estimated reliability or generalizability coefficients, and confidence in the adjustments will depend, in large part, on confidence in the estimates of the coefficients. The statistical adjustments will require accurate and precise estimates of coefficients or standard errors, based on generalizability studies that incorporate all sources of random variability and have adequate sample sizes; the adjustments need to meet the Goldilocks criterion—that the results are not being substantially underadjusted or overadjusted, but are just about right.

If one is going to correct for the bias associated with random errors in the prior test scores for VAM-based accountability programs, one will need to have accurate estimates of coefficients reflecting the overall error, including all significant sources of error. Getting accurate and precise estimates of all of variance components included in the error variance will require more careful attention to the estimation of the reliabilities of the prior test scores than has typically been the case in large-scale testing programs. Estimates of the variability due to the sampling of test tasks can be fairly precise because the estimates will generally involve a fairly large number of tasks, but estimates of the variability due to the sampling of occasions are not generally so precise because the estimates generally involve only two or three occasions. If estimates of reliability or generalizability coefficients or standard errors are to be used to adjust VAM scores in high-stakes contexts, it would be important to document the precision of the estimated errors and coefficients, for example, by estimating appropriate confidence intervals (Brennan, 2001b), as well as their accuracy in terms of the sources of error included.

As noted earlier, Boyd et al. (2013) have suggested a procedure for estimating the total error directly, and they suggested that the overall measurement error "is at least twice as large as that reported by the test vendor" (p. 629). It would be useful to compare estimates of the total error based on thorough generalizability studies to the results of analyses employed by Boyd et al. (2013); agreement between two disparate methods could increase confidence our estimates of the total error and associated coefficients.

## Conclusions

VAMs are designed to draw inferences about a school's or teacher's impact on student learning based on their students' test scores, while controlling for prior student learning (and possibly, other variables related to current performance). The intent is to obtain unbiased estimates of the teacher's or school's contribution to gains in student achievement.

Residual gain scores provide seemingly plausible adjustments for prior learning, and they circumvent the need for score scales that are linked across grades. The OLS regression models that are at the heart of residual gain estimates do not require that the scores on the current and prior tests be on the same scale; they do not even require that the current and prior tests measure the same variable. They do require that a linear model is plausible, and that the relationship is strong enough for the purpose at hand.

Any random errors in the prior scores used as an independent variable in the OLS regression of current scores on prior scores tend to add a positive bias to the residual gain scores for students with prior scores above the population mean, and they tend to add a negative bias to the residual gain scores for students with prior scores below the mean. The bias is associated with the well-known phenomenon of regression to the mean (Boggs, Spiegelman, Donaldson, & Schnabel, 1988; Deming, 1943; Fuller, 1987; Kane & Mroch, 2010), which does not cause problems in many contexts (particularly where prediction, as such, is the main goal), but if not adequately addressed, it does introduce bias into VAMs based on residual gain scores. The estimated VAM effects would tend to be overestimated for teachers and schools with relatively high-scoring students and they would tend to be underestimated for teachers and schools with relatively low-scoring students.

Note that if teacher assignments are based on the prior test scores to some extent, the bias due to unreliability in prior scores could be reduced (Lockwood & McCaffrey, 2014). This is not likely to be the case for assignments to schools, but could be an issue for assignments to teachers in schools.

Any source of bias in scores is troublesome, and a source of bias that has a substantial negative impact on VAM results for teachers and schools serving at-risk students could be especially problematic. As indicated above, the resulting biases in estimates of student gain scores could be quite large and positive for high-scoring students and could be quite large and negative for low-scoring students.

The differences in estimates of the teacher and school effects, which would involve averages over samples of students, would generally be far less dramatic, but given the way students are sorted into schools and classes in our educational system (Kalogrides & Loeb, 2013), the biases in the teacher and school effects are likely to be substantial in many cases. As indicated by the rough estimates presented earlier, if not corrected, the school-level bias could be a third of the overall, school-level true-score standard deviation. With a correction for the errors in the prior scores based on an estimate of internal consistency like coefficient alpha, this bias could be reduced substantially (by about 50% or more), but it would not be eliminated, unless the generalizability coefficient used for the correction takes account of all potentially substantial sources of error.

This source of bias can be controlled if good estimates of the generalizability/reliability coefficient reflecting all sources of random error in the prior scores is available, but such estimates are not routinely estimated for state tests. Error analyses that omit significant sources of random error tend to underestimate the error variance and overestimate the relevant generalizability coefficient, and adjustments that rely on such overestimates of the relevant coefficients will not adequately correct for the bias due to random errors in the prior scores.

## 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. https://doi.org/10.1086/508733

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4,* 351–383. https://doi.org/10.1162/edfp.2009.4.4.351

Berry, W. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage. https://doi.org/10.4135/9781412986427

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Boggs, P. T., Spiegelman, C. H., Donaldson, J. R., & Schnabel, R. B. (1988). A computational examination of orthogonal distance regression. *Journal of Econometrics, 38,* 169–201. https://doi.org/10.1016/0304-4076(88)90032-2

Boyd, D., Langford, H., Loeb, S., & Wyckoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics, 38,* 629–663. https://doi.org/10.3102/1076998613508584

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2007). *The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools* (Working Paper 10). Washington, DC: Center for Analysis of Longitudinal Data in Education Research.

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

Brennan, R. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *36*, 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x

Brennan, R. (2001b). *Generalizability theory*. New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0

Briggs, D. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50,* 204–226. https://doi.org/10.1111/jedm.12011

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

Briggs, D., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38,* 629–663. https://doi.org/10.3102/1076998613508317

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts.* New York, NY: Guilford Press.

Castellano, K., & Ho, A. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., Linn, R. L., Brennan R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399. https://doi.org/10.1177/0013164497057003001

Deming, W. E. (1943). *Statistical adjustment of data*. New York, NY: Dover.

Ferrara, S. (2006). Standardized assessment of individual achievement in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). Westport, CT: American Council on Education and Praeger.

Fuller, W. A. (1987). *Measurement error models*. New York, NY: Wiley. https://doi.org/10.1002/9780470316665

Haertel, E. (2013). *Reliability and validity of inferences about teachers based on student scores* (The 14th William H. Angoff Memorial Lecture). Princeton, NJ: Educational Testing Service.

Harris, C. (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher, 42,* 304–316. https://doi.org/10.3102/0013189X13495087

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement, 48,* 12–30. https://doi.org/10.1111/j.1745-3984.2010.00128.x

Kane, M., & Mroch, A. (2010). Modeling group differences in OLS and orthogonal regression: Implications for differential validity studies. *Applied Measurement in Education, 23,* 215–241. https://doi.org/10.1080/08957347.2010.485990

Kane, T., & Staiger, D. (2002). Volatility in school test scores: implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers in education policy* (pp. 235–269). Washington DC: Brookings Institution Press.

Kolen, M. (2006). Scaling and norming. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–220). Westport, CT: American Council on Education and Praeger.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.

Lockwood, J., & McCaffrey, D. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22–52. https://doi.org/10.3102/1076998613509405

McCaffrey, D., Castellano, K., & Lockwood, J. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice, 34*(1), 15–21. https://doi.org/10.1111/emip.12062

McCaffrey, D., Lockwood, J., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: Rand Corporation. https://doi.org/10.1037/e658712010-001

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29,* 67–101. https://doi.org/10.3102/10769986029001067

New York State Education Department. (2015). *The New York State Testing Program 2015: English language arts and mathematics grades 3–8.* Albany: New York State Education Department.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1–27. https://doi.org/10.14507/epaa.v18n23.2010

North Carolina Department of Public Instruction. (2014). *Reliability of the North Carolina end-of-grade and end-of-course assessments.* Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/eogeocreliabilities14.pdf

Rothstein, J. (2009). Student sorting and bias in value-added estimation: selection on observables and unobservables. *Education Finance and Policy, 3,* 537–571. https://doi.org/10.1162/edfp.2009.4.4.537

Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12,* 247–256. https://doi.org/10.1023/A:1008067210518

Sanford, E. (1996). *North Carolina end-of-grade tests: Reading comprehension, mathematics* (Technical Report No. 1). Raleigh: North Carolina Department of Public Instruction, Office of Instructional and Accountability Services.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1–26. https://doi.org/10.1207/S15324818AME1601_1

Winters, M., & Cowen, J. (2013). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42,* 330–337. https://doi.org/10.3102/0013189X13496145

Wisconsin Department of Public Instruction. (2015). *The Badger Exam 3-8: A Wisconsin Smarter Balanced Assessment Spring 2015 technical manual.* Madison: Wisconsin Department of Public Instruction.

## Suggested citation:

Kane, M. T. (2017). *Measurement error and bias in value-added models* (Research Report No. RR-17-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12153

**Action Editor:** James Carlson

**Reviewers:** Courtney Bell and J. R. Lockwood