An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education

S. Paul Wright, 10-MARCH-2010

Table of Contents

1. Introduction and Summary

Value-added modeling for educational evaluation is currently receiving considerable attention (Braun, Chudowsky and Koenig, 2010). As a result, a number of different value-added models (VAMs) have been proposed (Wainer, 2004). Two relatively recent additions to the VAM pantheon use nonparametric analyses that make fewer assumptions. These two models are the standardized gain model (SGM) and the student growth percentile model (SGPM). This paper investigates these two models, comparing them to some alternative models when applied to actual student test data. The SGM and the SGPM are described in Sections 2 and 3. Section 4 describes the data and additional models to which the SGM and SGPM are compared. Section 5 presents the results of the comparisons, and Section 6 discusses those results.

It will be seen that both the SGM and the SGPM, despite having the word "gain" in their designations, are regression models that calculate value-added by "averaging" regression residuals. Both models, therefore, inherit certain disadvantages that these regression models possess. First and foremost, regression coefficients are biased by the presence of measurement error in the predictor variable(s). This bias is particularly severe in single-predictor regression models (such as the SGM), but the bias can be ameliorated by inclusion of multiple predictors (as the SGPM does whenever possible). In value-added regression models, this bias tends to manifest itself by attenuating the regression coefficients so that the model behaves more like a "status" model (which measures how much students know) as opposed to a "value-added" model (which measures how much students have learned). The result is that school and/or teacher effectiveness, as estimated by these models, tends to be correlated with school/classroom composition, with high-poverty/high-minority schools/classrooms being more likely to be evaluated as ineffective. The results discussed below demonstrate this.

A second disadvantage that these two models share is due to their nonparametric methodology. The *advantage* of nonparametric models is that they make fewer assumptions and are therefore applicable in a wider variety of situations. Assumptions typically made in parametric regression models, but not made in nonparametric models, include normality, linearity, and homoscedasticity (constant variance). However, if these parametric assumptions are, in fact, true (at least approximately), then estimates from the parametric model are more precise than those from a nonparametric model. (If the assumptions are not true, the nonparametric model is demonstrably better.) The price that is paid for the additional generality of a nonparametric model is a loss of precision, that is, more uncertainty about which teachers or schools are being effective or ineffective. The results discussed below suggest that this additional uncertainty particularly affects the SGPM.

2. The Standardized Gain Model (SGM)

Reback (2008) is an often cited reference for the SGM. Variants of this approach have been used in Tennessee (performanceincentives.org/research/mnps.asp) and in Texas (Springer, et al., 2009). McCaffrey, Han, and Lockwood (2010) include this approach in their comparisons of a large number of value-added models; they refer to it as the "lookup table" approach.

Let:
Y = current (post-test) score, e.g., 2009 6th grade math score.
X = previous (pre-test) score, e.g., 2008 5th grade math score.
G = Y − X = gain.
Mean[G|X] = mean of G values for all students with a particular X value.
SD[G|X] = standard deviation of G values at a particular X value.
Z = {G − Mean[G|X]} / SD[G|X].

"Z" is the standardized gain. In value-added applications, the standardized gain typically would be averaged for all students of a particular teacher, school or district to obtain an indicator of effectiveness. (Reback, 2008, and Ladd and Lauen, 2009, use the standardized gain as the dependent variable in further analyses that are not relevant in the present paper.) To see the connection to traditional regression analysis, note the following.

Mean[G|X] = Mean[(Y−X)|X] = Mean[Y|X] − X.
SD[G|X] = SD[(Y−X)|X] = SD[Y|X].
Z = (Y − Mean[Y|X]) / SD[Y|X].

Z, the standardized gain, is seen to be a standardized regression residual since Mean[Y|X], the conditional mean of Y given X, is the line (or curve) of the regression of Y on X and SD[Y|X], the conditional standard deviation, is the square root of the residual variance in such a regression. In a traditional linear regression, Mean[Y|X] would be fitted as a straight line (the assumption of linearity) and SD[Y|X] would be the same for every X (the assumption of constant variance or homoscedasticity). The SGM relaxes these assumptions, allowing Mean[Y|X] to follow a nonlinear relationship and allowing for different variances at different values of X.

One difficulty in implementing the SGM is that for some X values there may be very few Y values. This can result in wild swings in the values of Mean[Y|X] and SD[Y|X]. In fact, in the examples below, some X values had only one associated Y value, resulting in a standard deviation of zero, making the calculation of a standardized gain impossible. This difficulty was addressed in this paper by smoothing Mean[Y|X] and SD[Y|X], as functions of X, using local regression (the LOESS procedure in SAS/STAT® 9.2).

3. The Student Growth Percentiles Model (SGPM)

Betebenner (2009a, 2009b) is the primary advocate of this modeling approach, which has been implemented in Colorado (cde.state.co.us/research/GrowthModel.htm: follow the "Documentation" and "Resources for Users" links for more information), Massachusetts (doe.mass.edu/mcas/growth), and Indiana (doe.in.gov/growthmodel). This approach is closely related to the SGM. Recall that in the SGM, for each unique X value, there exists a collection of Y values from which conditional statistics are calculated, namely the conditional mean and standard deviation (Mean[Y|X] and SD[Y|X]). Other conditional statistics could be calculated as well, such as medians, quartiles, etc. In particular, for each Y value in the set of Y values at a given X value, one could calculate the conditional percentile rank within that set of Y values. This conditional percentile rank is the student growth percentile (SGP). It indicates a student's level of achievement on test Y compared to other students who had the same prior test score X. It may be helpful to think about the SGP in terms of the standardized gains in the SGM of Section 2. Let Z be the standardized gain (i.e., the standardized regression residual) for a particular student. If the conditional distribution is normally distributed (as would be assumed in a parametric regression), then the SGP (the conditional percentile rank) is

$$SGP = 100 \, \Phi(Z)$$

where Φ is the standard normal cumulative distribution function that evaluates the area to the left of Z under the standard normal "bell" curve.

The actual calculation of SGPs in the SGPM is done differently. The problem of instability due to small numbers of Y values at some X values, which was dealt with by smoothing in the SGM, is much more of a problem when calculating SGPs. This problem is compounded by the fact that, in the SGPM (unlike the SGM), multiple prior test scores are used when available. The result is that each set of Y values for calculating conditional percentiles is for students with the exact same set of prior test scores on multiple tests. Most of these sets of Y values will be quite small. Some kind of smoothing is therefore essential. The smoothing in the SGPM is accomplished by a statistical modeling technique called quantile regression. This differs from standard multiple regression in several ways. One, there is no assumption of normality. Two, there is no assumption of homoscedasticity. Three, the criterion for fitting is not minimizing squared errors but minimizing absolute errors. This provides a measure of robustness against outliers in the response variable; however, it does not protect against outliers in the predictor variables. Quantile regression models, like traditional multiple regression models, are "linear" models; but just as in traditional regression, nonlinear relationships can be accommodated, for example, by including polynomial terms in the regression. In the SGPM, nonlinearity is incorporated by modeling each prior test score using a smooth curve (a cubic B-spline with 4 knots). In quantile regression, a separate regression equation is estimated for each desired quantile; in the SGPM there are 99 desired quantiles corresponding to percentiles 1 to 99. A student's actual 2009 math score is compared to the fitted value from each of the 99 regressions; the percentile that gives the best match to the student's actual score is the student's SGP.

For the purpose of obtaining a value-added indicator of effectiveness the SGPs, which are calculated for each student, must be summarized at the desired level of aggregation. In those places where the SGPM is currently in use, the aggregation is primarily at the school level. This paper focuses on teacher-level assessment. For the SGM of Section 2, the summary statistic was the arithmetic mean. Since percentiles, and thus SGPs, do not possess an interval-scaled metric, the arithmetic mean is not appropriate. The median is used instead.

4. Data and Models for Comparison

The SGM and the SGPM, along with several other VAMs for comparison, were applied to three sets of statewide test scores. Specifically, these were the 5th grade, 6th grade and 7th grade spring 2009 math scores on the state criterion-referenced end-of-grade tests (the tests used for NCLB) from a state served by SAS® EVAAS® for K-12. Math test scores from up to three prior years/grades were used as predictors in the SGPM; the SGM uses only one prior score.

For value-added, aggregation was done at the teacher level. Note that, especially in grades 6 and 7, a teacher may have taught more than one class. Students from all classes were combined to obtain a single teacher value-added indicator. A student was included in the analyses only if a single teacher claimed at least 80% responsibility for that student in mathematics in 2009. Aggregate statistics (means, medians) were calculated only for teachers who had at least 8 students available for aggregation (i.e., 8 SGPs, 8 standardized gains). For the SGM, aggregation was done by calculating the mean of the standardized gains. For the SGPM, aggregation was done by calculating the median of the SGPs. The following additional aggregate measures were obtained for comparison.

Additional SGPMs. In the standard SGPM, an SGP is obtained for each student who has at least one prior math test score (in the previous grade in the previous year, excluding students who did not progress normally), but it uses multiple prior math scores when available (up to 3 years in this paper). Thus, the SGP is based on three prior math scores for some students (from years 2008, 2007, 2006), on two prior math scores for other students (2008, 2007), and on one prior math score for other students (2008). In fact, most students had three prior test scores (except 5th graders for whom only two prior scores are available from 4th and 3rd grades). For comparison (especially for comparison to the SGM, which uses only one prior score), SGPs were also obtained using only one prior math score (2008) and only one or two prior math scores (2008 and 2007).

Mean score and mean gain. The mean 2009 math score (in grade 5, 6 or 7) was calculated for each teacher's students. This, of course, is not a value-added measure, but it provides a useful reference against which to compare the value-added model results. Similarly, the mean 2009 math gain was calculated for each teacher (2009 5th grade math minus 2008 4th grade math, 2009 6th grade math minus 2008 5th grade math, or 2009 7th grade math minus 2008 6th grade math).

SAS EVAAS estimated gain. Since the data used in the analyses is from a state served by SAS EVAAS, it was possible to obtain estimated teacher gains from the EVAAS layered teacher model. This is a multivariate, longitudinal mixed model that uses up to five years of scores in four subjects (math, reading/language arts, science, social studies) for each student. For more information about this model see Sanders, et al. (1997) or McCaffrey, et al. (2004). As mentioned above, only students who were claimed (with at least 80% responsibility) by a single teacher were included in the analyses for this paper. Because the EVAAS gains were obtained from SAS EVAAS, it was not possible to apply this restriction to the EVAAS model. The EVAAS model accommodates team teaching and departmentalized instruction so that all students are included. Therefore, there were students included in the EVAAS gain calculations who were not included in any of the other models. However, for the purpose of making model comparisons, EVAAS gains were included only for those teachers for whom a mean score was available.

ANCOVA model. The analysis of covariance model used the 2009 math score as the response, the teacher identifier as the classification variable, and up to 12 prior test scores as covariates (up to 3 years in 4 subjects except 5th grade for which only two prior grades were available). The teacher effect was treated as a random effect (that is, the teacher effects are shrinkage, or empirical Bayes, estimates). To accommodate students with fewer than all 12 covariates (all 8 for 5th grade), a "composite prior score" (or "projected score") was calculated using whichever of the covariates was available for each student. The method of obtaining these projected scores is described in Wright, et al. (2006). The predicted/composite score was then used as the covariate in the analysis of covariance. In order to minimize bias due to measurement error in the covariates, only composites based on at least 3 prior test scores were used.

Additional ANCOVA models. For comparison, additional ANCOVA models were run using just one prior math score (from 2008), using two prior math scores (2008 and 2007), and using three prior math scores (2008, 2007, 2006). That is, these three models used the same prior scores as were used in the SGP models. However, unlike the above ANCOVA model and unlike the SGP models, but in conformity to the usual practice with ANCOVA modeling, students having fewer than the full set of covariates were excluded from the analyses.

Calculations. The SGPs were calculated using the SGP package (version 0.0-4) in R version 2.9.2 (R Development Core Team, 2009). SAS EVAAS teacher gains were obtained from SAS EVAAS modeling software. All other results were obtained using SAS software version 9.2.

Calculation of standard errors. In statistics, it is standard practice to provide a measure of uncertainty, e.g., a standard error, for estimated quantities. This was done for the estimates of teacher effectiveness from each of the above models. For the EVAAS Gains and the ANCOVA models, standard errors are supplied routinely by the software. For the SGM, the usual standard error of the mean was calculated for each teacher's students when the mean was calculated. For the SGPM, the Colorado model calculates a "comparison region" around a median SGP as $\pm 3 \ast IQR / \sqrt{N}$ where IQR is the interquartile range of the SGPs and N is the number of students. (See cde.state.co.us/cdeassess/sar_info.html, the "Academic Growth Calculation Example".) This was re-expressed as a standard error as follows.

For normally distributed data, the standard error of the median is 1.2533 $\sigma/\sqrt{N}$ where σ is the population standard deviation (Stuart and Ord, 1994, §10.10). Again assuming a normal distribution, σ can be estimated as IQR/1.3490 (Stuart and Ord, 1994, §10.11). This gives an estimated standard error of the median as 0.929*IQR/$\sqrt{N}$.

5. Results

A primary concern among educators about any evaluation system, value-added or otherwise, is fairness. In particular, teachers in high-poverty/high-minority classrooms fear that they may be at a disadvantage compared to teachers in low-poverty/low-minority classrooms. One way to address this concern is to look at the relationship between classroom poverty/minority composition and the evaluation criteria. Table 2 shows the correlations between the criteria from each of the models discussed in Section 4 (listed for convenience in Table 1) and a teacher-level aggregate measure of poverty: the percentage of each teacher's students who were eligible for free-or-reduced-price lunch (%FRPL). (Similar correlations were obtained using percent minority rather than percent FRPL. The pattern of these correlations was similar to Table 2, but the correlations tended to be much closer to zero.)

Mean Score	Arithmetic average of student scores.
Mean Gain	Arithmetic average of student gains (2009 math score minus 2008 math score in previous grade).
EVAAS Gain	Estimated gain from "layered" EVAAS multivariate, longitudinal mixed model.
ANCOVA: ≥ 3 covariates	Analysis of covariance with random teacher effects using up to 12 covariates (all available scores in four subjects from three previous years) with a minimum of 3 covariates.
ANCOVA: 3 math	Random effects analysis of covariance using math scores from three previous years as covariates.
ANCOVA: 2 math	Random effects analysis of covariance using math scores from two previous years as covariates.
ANCOVA: 1 math	Random effects analysis of covariance using a single math score from the previous year as the covariate.
Mean Standardized Gain	Arithmetic average of "standardized gains" (standardized regression residuals) as described in Section 2.
Median SGP: ≤ 3 math	Median of student gain percentiles from quantile regression using math scores from one, two, or three previous years (whichever is largest) as described in Section 3.
Median SGP: ≤ 2 math	Median of student gain percentiles from quantile regression using math scores from one or two previous years.
Median SGP: 1 math	Median of student gain percentiles from quantile regression using math scores from the previous year.

Table 1. Descriptions of evaluation criteria used in comparisons of Section 5 and Table 2.

	Grade 5		Grade 6		Grade 7	
	N	Corr	N	Corr	N	Corr
Mean Score	2322	−0.66	1085	−0.69	881	−0.70
Mean Gain	2319	−0.08	1082	+0.03	880	+0.05
EVAAS Gain	2322	−0.14	1085	−0.03	881	−0.03
ANCOVA: ≥ 3 covariates	2318	−0.13	1083	−0.07	880	−0.08
ANCOVA: 3 math			1074	−0.16	875	−0.17
ANCOVA: 2 math	2300	−0.28	1080	−0.20	877	−0.21
ANCOVA: 1 math	2319	−0.38	1082	−0.27	880	−0.32
Mean Standardized Gain	2319	−0.34	1082	−0.24	880	−0.27
Median SGP: ≤ 3 math			1082	−0.13	880	−0.16
Median SGP: ≤ 2 math	2319	−0.23	1082	−0.16	880	−0.18
Median SGP: 1 math	2319	−0.31	1082	−0.21	880	−0.24

Table 2. Correlations between teacher effectiveness indicators from various models and classroom composition. N is the number of teachers having at least 8 students available for use in the model. "Corr" is the correlation between estimated teacher effectiveness and percent free-and-reduced-price-lunch eligibility (%FRPL).

The correlations of %FRPL with average scores ("Mean Score" in Table 2) demonstrate why educators are concerned. The large negative correlations indicate that teachers in high-poverty environments will tend to get poorer evaluations when evaluated by student "status." One of the attractions of value-added modeling is its potential to avoid the unfairness of status criteria.

This potential to avoid overt unfairness is borne out by most of the remaining criteria in Table 2, all of which (other than "Mean Score") are generally considered to be value-added criteria. For the first three VAMs ("Mean Gain", "EVAAS Gain", "ANCOVA: ≥ 3 covariates"), correlations with %FRPL are relatively small (close to zero in grades 6 and 7). For the remaining VAMs, all of which involve regressions with a limited number of predictor variables (3 or fewer), the correlations become noticeably larger in magnitude. All of the correlations are highly statistically significant (p-value < 0.0001) except 6th and 7th grade "Mean Gain" and "EVAAS Gain" (p-values from .14 to .44) and 6th and 7th grade "ANCOVA: ≥ 3 covariates" (p-values .03 and .02, respectively, i.e., significant but not highly significant). However, given the large sample sizes, statistical significance is a poor indicator of the practical importance of the correlations.

6. Discussion: Threats to Value-Added Modeling

The attention that value-added modeling has received is due to its promise of a fairer, more objective assessment of the extent to which teachers and schools are helping their students learn. The results in Table 2 suggest that, by one measure of fairness (correlation with a poverty indicator), value-added models succeed at this, though some VAMs are more successful than others. However, as critics of value-added modeling are happy to point out, there are a number of issues that threaten to prevent value-added modeling from delivering on its promise. Several of these threats are discussed below with particular emphasis on how they affect the SGM and SGPM.

Measurement error. The effect of measurement error in the predictor variable in a one-predictor linear regression model is well-known: The regression coefficient is biased toward zero. In the context of value-added indicators based on "average" (mean or median) regression residuals, the consequence is that the "value-added" effectiveness indicator is more like a "status" indicator. (In the extreme case of a regression coefficient biased completely to zero, the regression residuals are simply deviations from a grand mean and the averaged residuals are equivalent to calculating the mean score.) This measurement-error-induced bias can be ameliorated by including multiple predictors in the regression so that their measurement errors tend to average out.

To the extent that a value-added indicator measures status rather than value-added, it inherits the unfairness of a status indicator, e.g., it tends to be negatively correlated with classroom poverty. This is well-displayed in Table 2. The single-predictor regression models (SGM, SGPM and ANCOVA with one predictor) have correlations with %FRPL ranging from –0.21 to –0.38. With the two-predictor SGPM and ANCOVA, the correlations become smaller; with 3 predictors they are smaller still. With the general ANCOVA, in which many students have 8 predictors (5th grade) or 12 predictors (6th and 7th grade), the correlations are even lower, approaching zero in grades 6 and 7.

For models that estimate gain (simple mean gain, EVAAS gain), the effect of measurement error is not so much to increase bias but to increase estimation error. (However, there are other ways these estimates can become biased; see the section on missing data below.) As a result, they tend to have small correlations with %FRPL, near zero in grades 6 and 7.

Estimation error. Figures 1 and 2 each contain a scatterplot of t-values from two value-added models. Each t-value has the form of an "effect" divided by its standard error. For this discussion, an "effect" is a value-added indicator for a particular teacher minus the average value-added indicator for all teachers; that is, the effect for an average teacher is set to zero. Standard errors were discussed at the end of Section 4. Dividing an effect by its standard error puts it on a familiar scale. In conventional statistical terminology, t-values greater than 2 in magnitude are "statistically significantly different from zero." Teachers with t-values greater than +2 could be considered "demonstrably better than average" in the sense that their students made demonstrably more academic progress than students of the average teacher. Those with t-values below –2 are "demonstrably poorer than average." (In Figures 1 and 2 there are vertical and horizontal reference lines at ±2 as well as at zero.)

Figure 1 plots t-values from the SGM against the EVAAS gain model. Plots of most other models against one another have a similar appearance. The correlation is high (generally about 0.90); and while there is considerable agreement about which teachers are above average, below average, or not detectably different from average, there is enough disagreement to make the choice of a value-added model an important decision. The one model that stands out in such scatterplots is the SGPM, shown in Figure 2 plotted against the EVAAS gain model. This plot shows noticeably more scatter especially at the high and low ends of the scale. This suggests that there is more instability in the estimates from the SGPM than from other models.
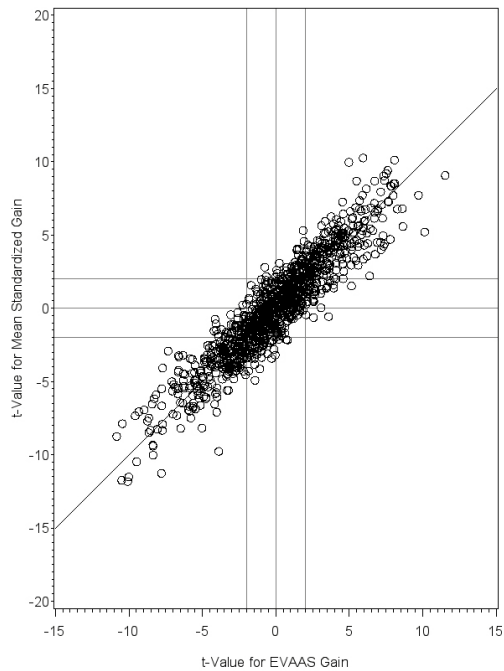
Figure 1. Scatterplot of 6th grade t-values from SGM versus t-values from EVAAS model with diagonal 1-to-1 line and reference lines at zero and ±2.

Greater instability in the SGPM estimates is not unexpected. Estimates from nonparametric methods are, by their nature, less stable than those from parametric analyses. However, the SGPM estimates seem even more unstable than the SGM estimates, which are also nonparametric. The extra instability in the SGPM may be due, in part, to the quantile regression algorithm. Estimating a regression at extreme percentiles (near zero or 100) can be challenging even with a large data set. The task is made more challenging in the SGPM by the use of nonlinear relationships. The SGPM with 3 prior scores, by using a cubic spline with 4 knots for each prior score, is actually estimating 21 regression coefficients (in addition to an intercept). With over 50,000 students in the regression (see below), this is not an unreasonable task for percentiles near the center of the distribution (e.g., 50th percentile); but at the extremes, the estimated regression coefficients are likely to contain a large amount of estimation error. The number of students whose SGPs are in the extremes is non-trivial. SGPs, being percentiles, are approximately uniformly distributed. This means, for example, that 20% of the students have SGPs ≤10 or ≥90. Of course, the teacher-level median SGPs are less extreme, and less unstable, than the individual student SGPs; for the analyses in this paper, the median SGPs were mostly between 25 and 75. However, the estimated standard errors of these median SGPs, which are based on sample interquartile ranges, are likely to be quite unstable, especially in small classrooms and especially if those small classrooms have many students with extreme SGPs. Since these standard errors are necessary to discriminate reliably among "good" teachers, "bad" teachers, and "average" teachers, this instability is a serious threat to the usefulness of the SGPM. See the section on "Measurement error and nonlinearity" for additional comments on the SGPM.

Missing data. As mentioned above, one solution to the problem of measurement error in regression models is to include more predictor variables in the regression. The flaw in this plan is that students miss tests. The more predictors there are, the greater the chance that a student will have missing data for at least one predictor; and in traditional regression modeling, a student must have data on all predictors (as well as the response) to be included in the analysis. The impact of missing data is shown in Table 3, which indicates how many students were used in each model described in Sections 2 through 4. Nearly 95% of the students had a prior math score: These students were used in the "Mean Gain" calculation as well as the SGM, all the SGPMs, and the ANCOVA with one prior math score. In the ANCOVA with 2 prior scores, the percentage falls to about 88%; with 3 prior scores it falls to about 83%. This also affects the number of teachers who can be evaluated (shown in Table 2) since, for example, a teacher who has 8 students with scores may have fewer than 8 students with gains. Notice that the number of students used in the "EVAAS Gain" is 5-6% larger than for "Mean Score" (even though the same teachers were used for both). As noted in Section 4, the EVAAS model includes students who were taught by multiple teachers while the other models were limited to students taught primarily by a single teacher. It would be possible to modify the other models to handle multiple teachers per student, but in practice this has not yet been done.

	Grade 5		Grade 6		Grade 7	
	Students	Percent	Students	Percent	Students	Percent
Mean Score	57990	100	56579	100	56060	100
Mean Gain	54887	94.6	53272	94.2	52982	94.5
EVAAS Gain	60904	105.0	60309	106.6	59513	106.2
ANCOVA: ≥ 3 covariates	55427	95.6	54108	95.6	53721	95.8
ANCOVA: 2 math	51022	87.8	50064	88.5	49625	88.5
ANCOVA: 3 math	46548	82.3	46546	83.0		

Table 3. Number of students used in each model. Numbers for one-predictor ANCOVA, SGM, and SGPM are the same as for "Mean Gain." See the text for an explanation of the EVAAS Gain model.

Even with only 83% of the students (the worst case in Table 3), that still leaves nearly 50,000 students, so why is missing data a problem? One problem is that each teacher does not have 50,000 students. It is the number of students per teacher that is relevant when estimating teacher effectiveness, so the loss of even a handful of students can have a big impact on the stability (standard error) of an estimate. This is particularly true in elementary school where most teachers teach a single class (the mean number of students per teacher who had a "Mean Score" in 5th grade was 25). It is less of a concern in middle school (6th grade mean n=52, 7th grade mean n=64).

A second problem is bias. With the exception of the EVAAS model, the estimates from all the models assume students are "missing completely at random." Roughly speaking, this means that the available (non-missing) scores are a fair representation of the missing scores. This is demonstrably not the case; rather, scores near the low end of the distribution are more likely to be missing than high-end scores. Such non-random missingness can produce biased estimates even for "Mean Gain" estimates (Wright, 2004). One of the attractive features of multivariate, longitudinal models such as the EVAAS model is its ability to minimize such bias (Wright, 2004; Lockwood and McCaffrey, 2007, p. 29). This bias reduction results from including all available student scores and from using estimation methodology (maximum likelihood), which only requires "missing at random" rather than "missing completely at random." (Roughly speaking, "missing at random" means that the missing scores are "predictable" from the non-missing scores.)

Additional topics: Measurement error and nonlinearity. Betebenner (2009b) demonstrates the need to allow for nonlinearity and heteroscedasticity in the SGPM by showing graphical results from two one-predictor quantile regressions (his Figure 2: grade 6 math regressed on grade 5 math). One of the quantile regressions uses linear relationships; the other allows for nonlinearity using splines as in the SGPM. The spline regressions clearly show an S-shaped nonlinear pattern with some heteroscedasticity in the tails, particularly the lower tail (low 5th grade math scores). It should be noted that one of the characteristics of standardized tests is that there is more measurement error in high and low scores than in scores near the center of the distribution. This suggests that the nonlinearity and heteroscedasticity seen in Betebenner's (2009b) Figure 2 is the result of measurement error. This was confirmed by doing a simple simulation. Bivariate normal X and Y scores were simulated so that the true regression of Y on X would be linear and homoscedastic. Y was then regressed on X using quantile regression with splines. The resulting regressions appeared linear and homoscedastic, as expected. Then random error was added to X and Y with the amount of error increasing toward the extremes. Again Y was regressed on X using quantile regression with splines. The results showed the characteristic S-shaped curve with heteroscedasticity. This suggests that the SGPM regressions are being biased in multiple ways by the presence of measurement error in the predictors. Not only is the overall relationship biased toward zero (toward a status model), but this bias is exacerbated for low and high scoring students (and thus for teachers of predominantly low or high scoring students) by modeling measurement-error-induced nonlinearity.
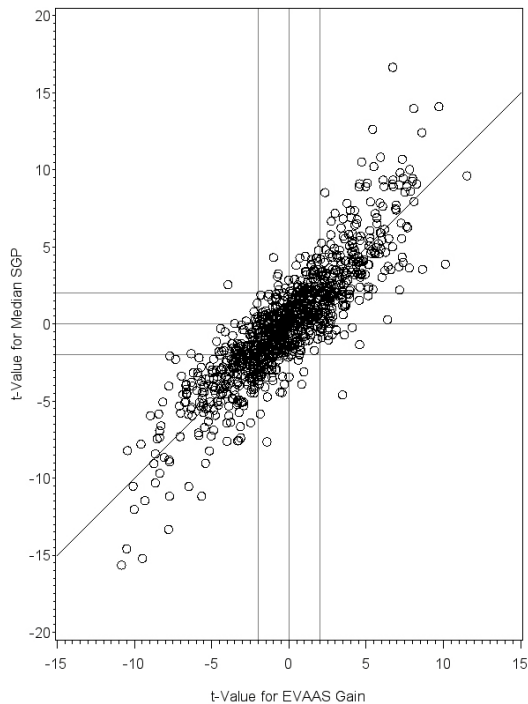
Figure 2. Scatterplot of 6th grade t-values from SGPM using up to 3 prior math scores versus t-values from EVAAS model with diagonal 1-to-1 line and reference lines at zero and ±2.

Additional topics: Analysis of Gains. The purpose of this paper was to evaluate the SGM and SGPM approaches to value-added assessment, not to conduct a comprehensive comparison of all possible value-added models. For that purpose, see McCaffrey, et al. (2004, 2010). However, the observant reader cannot fail to have noticed that, at least in terms of "fairness" as indicated by Table 2, the very simple, easily calculated, easily understood "Mean Gain" indicator of value-added seems to be competitive with the statistically sophisticated, computationally demanding "EVAAS Gain" indicator. A word of warning is therefore in order for those who are tempted to conclude that simple gains are all that is needed for value-added analysis. First, gain scores are relatively noisy. As noted above, the effect of measurement error on "Mean Gain" estimates is not to produce bias but to increase estimation error. McCaffrey, et al. (2010, p. 133) have made note of this: "Gain-score methods are like a coin flip – fair but capricious." Second, there is the problem of missing data. This exacerbates the noisiness of gain-scores by reducing the sample size, and it also has the potential to produce bias due to non-random missing data. In fairness, it should be noted that if there is bias in the "Mean Gain" estimates, the impact of that bias is not to produce the kind of obvious unfairness that produces negative correlations with poverty as seen for many of the estimates in Table 2. To summarize, the major disadvantage of using gain scores is their instability.

Additional topics: "Total" versus "pooled-within" regression. Of the models compared in this paper, the ANCOVA models, the SGPMs, and the SGM (all except "Mean Score," "Mean Gain," and "EVAAS Gain") may be characterized as regression models in which a student's current-year score is predicted from one or more prior-year scores. However, these models use two rather different philosophies for estimating regression coefficients and, consequently, for obtaining regression residuals from which value-added effectiveness indicators are obtained. The ANCOVA models estimate "pooled-within-teacher" regression coefficients that take into account the grouping of students within teachers. In contrast, the SGM and SGPMs calculate "total" regression coefficients that ignore the grouping of students within teachers. Mathematically, in ordinary regression, the "total" coefficients can be shown to be a weighted average of the "pooled-within" coefficients and the "between" regression coefficients, where the "between" coefficients are obtained by regressing teacher-level mean current-year scores on teacher-level mean prior-year scores (using the number of students per teacher as weights in the regression). If the "between" regression is different from the "pooled-within" regression, then the total regression model estimates (SGM, SGPM) will differ from pooled-within model estimates (ANCOVA). It seems unlikely that these regressions would differ much when using statewide data (although this was not investigated for this paper and merits future investigation). It would be of greater concern if the SGM or SGPM were applied on a smaller scale, for example, within a single district.

7. Conclusions

The primary purpose of this paper was to examine two relatively recent value-added models, the standardized gain model (SGM) and the student growth percentile model (SGPM), and to assess their potential usefulness for value-added evaluation compared to other value-added models. It was shown that both of these models are a type of regression model that estimates value-added effectiveness by "averaging" regression residuals. These models differ from other regression models used in value-added modeling (analysis of covariance, hierarchical linear models) in taking a non-parametric approach to the regression, thereby avoiding the traditional regression assumptions of normality, linearity, and homoscedasticity (constant variance). Nevertheless, these models inherit the disadvantages that are common to all regression models. The most serious disadvantage is that measurement error in the predictor variable(s) produces biased estimates. This bias shows itself in estimates that are correlated with school/classroom composition, with high-poverty/high-minority schools/classrooms being more likely to be evaluated as ineffective. This bias is particularly severe in the SGM, which uses only one predictor variable. The bias can be reduced by using multiple predictor variables in the regression; the SGPM (but not the SGM) does this whenever possible. However, among the models compared in this paper, the SGPM produced the most unstable estimates. This extra instability appears to be the result of using quantile regression methodology combined with nonlinear relationships (cubic splines).

Like all value-added models, the SGM and the SGPM have advantages and disadvantages that should be carefully considered by potential users. In the case of the SGM, the only apparent advantage is its transparency – it is easy to explain. This is more than outweighed by its bias against teachers of socioeconomically disadvantaged students caused by measurement error in the test scores. The SGPM potentially suffers the same disadvantage (without the benefit of transparency) unless it incorporates multiple prior test scores (as it does whenever possible). In addition, the SGPM suffers more from instability than do most other models, reducing its ability to discriminate reliably among teachers of below average, average, and above average effectiveness.

References

Braun, H., Chudowsky, N., and Koenig, J. A., editors (2010). *Getting Value Out of Value-Added: Report of a Workshop.* Washington, D.C.: The National Academies Press. Available at nap.edu/catalog/12820.html.

Betebenner, D. W. (2009a). Norm- and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practices*, Vol. 28, no. 4, pp. 42-51. A pre-publication version is available at nciea.org/publications/normative_criterion_growth_DB08.pdf.

Betebenner, D. W. (2009b). Growth, Standards and Accountability. National Center for the Improvement of Educational Assessment. Accessed 3-1-2010 at nciea.org/publications/growthandStandard_DB09.pdf.

Ladd, H. F., and Lauen, D. L. Status vs. Growth: The Distributional Effects of School Accountability Policies. *Journal of Policy Analysis and Management* (2009, forthcoming). CALDER working paper 21 at caldercenter.org/publications/calder-working-paper-21.cfm.

Lockwood, J. R., and McCaffrey, D. F. (2007). Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement. *Electronic Journal of Statistics*, Vol. 1, pp. 223-252.

McCaffrey, D. F., Han, B. and Lockwood, J. R. (2010). Turning Student Test Scores into Teacher Compensations Systems. Chapter 6 in Springer, M. G. (ed.), *Performance Incentives: Their Growing Impact on American K-12 Education*. Washington, D. C.: Brookings Institution Press.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. J*ournal of Educational and Behavioral Statistics*, Vol. 29, No. 1, pp. 67-101.

R Development Core Team. (2009). R: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from R-project.org.

Reback, R. (2008). Teaching to the Rating: School Accountability and the Distribution of Student Achievement. *Journal of Public Economics*, 2008, Vol. 92, pp. 1394-1415.

Sanders, W. L., Saxton, A. M., and Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Educational Measure?* pp. 137-162. Thousand Oaks, CA: Mean Score

Springer, M. G., Podgursky, M. J., Lewis, J. L., Ehlert, M. W., Gronberg, T. J., Hamilton, L. S., Jansen, D. W., Lopez, O. S., Peng, A., Stecher, B. M, & Taylor, L. L. (2009). Texas Educator Excellence Grant (TEEG) Program: Year Three Evaluation Report. Austin, TX: Texas Education Agency. Accessed 12-24-2009 at ritter.tea.state.tx.us/opge/progeval/TeacherIncentive/TEEG_Y3_0809.pdf.

Stuart, A, and Ord, J K. (1994). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, Sixth Edition. London: Edward Arnold.

Wainer, H., editor (2004). Value-Added Assessment Special Issue. *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1.

Wright, S. P. (2008). Estimating Educational Effects using Analysis of Covariance with Measurement Error. Paper presented at CREATE/NEI Conference, Wilmington, NC, October 2008. Accessed 12-24-2009 at createconference.org/documents/archive/2008/2008wright.pdf.

Wright, S. P. (2004). Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment Without Imputation. Paper presented at CREATE/NEI Conference, Colorado Springs, CO, July 2004. Accessed 12-24-2009 at createconference.org/documents/archive/2004/Wright-NEl04.pdf.

Wright, S. P., Sanders, W. L., and Rivers, J. C. (2006). Measurement of Academic Growth of Individual Students toward Variable and Meaningful Academic Standards. Pages 385-406 in R. Lissitz (Ed.), *Longitudinal and Value Added Models of Student Performance*. Maple Grove, MN: JAM Press. Available online at sas.com/govedu/edu/wrightsandersrivers.pdf.

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 60,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.