Simple Choices among Aggregate-Level Conditional Status Metrics:
From Median Student Growth Percentiles to Value-Added Models

Katherine E. Castellano
University of California, Berkeley
kcastel@berkeley.edu
240-447-6907

Andrew D. Ho
Harvard Graduate School of Education
Andrew_Ho@gse.harvard.edu
617-496-2408

Author Note
KATHERINE CASTELLANO is an Institute of Education Sciences postdoctoral fellow at the University of California, Berkeley, 4419 Tolman Hall, Berkeley, CA 94720; email: kcastel@berkeley.edu. Her research interests involve the application of statistical models to educational policy issues, such as the use of the growth models in accountability systems.

ANDREW HO is an Assistant Professor at the Harvard Graduate School of Education, 455 Gutman Library, 6 Appian Way, Cambridge, MA 02138; email: Andrew_Ho@gse.harvard.edu. His research interests involve educational accountability metrics that incorporate inferences about proficiency, growth, college readiness, and value added.
Abstract

Aggregate-Level Conditional Status Metrics (ACSMs) describe the status of a group by referencing current performance to expectations given past scores. This paper focuses on seven ACSMs that condition only on past scores, including median Student Growth Percentiles (Betebenner, 2009), aggregated Percentile Ranks of Residuals (Castellano & Ho, 2012) and covariate-adjustment “value-added” models (e.g., McCaffrey et al., 2004). The authors define “simple” choices among these ACSMs (e.g., between mean- and median-based aggregation functions, or between fixed and random effects). They evaluate the impact of these choices upon three practical factors: school rankings, robustness across inclusion of prior-year data, and invariance to nonlinear scale transformations. Findings include expressions of practical differences between ACSMs and suggest considerable advantages of mean-based over median-based metrics.

Keywords: Student Growth Percentiles, Value-Added Models, group-level growth, conditional status
Simple Choices among Aggregate-Level Conditional Status Metrics:

From Median Student Growth Percentiles to Value-Added Models

Recent educational policies have expanded the focus of accountability measures from student proficiency at a single time point to student score histories over time. This paper reviews metrics that use these longitudinal data to describe the status of groups of students in terms of empirical expectations given past scores. Referencing status to expectations is intuitively appealing. A group of students may have low achievement but higher achievement than expected given their past scores. Group performance may be better interpreted and improved in light of these expectations.

These methods use regression models that locate individuals and groups in empirically “comparable” reference groups based on their past scores. This may be described as a kind of norm-referencing. Following Castellano and Ho (2012), we call these metrics “conditional status metrics” (CSMs), because they frame current status in terms of conditional distributions given past scores. This class of metrics includes residuals from ordinary least squares (OLS) regression models (i.e., residual gain scores, Manning & DuBois, 1962). Another CSM is quantile-regression-based Student Growth Percentile metric (SGPs; Betebenner, 2008a, 2009) that is in active or preliminary use in 25 states (Betebenner, 2010a).

Individual-level CSMs can serve multiple purposes from student-level classification, accountability, and selection to enhancing student score reports for student, parent, and teacher audiences. However, large-scale decisions and policy-relevant inferences are often supported by aggregate-level metrics that describe groups of students instead of individual students. Current educational policies require longitudinal data summaries at an aggregate level, such as

The popularity and utility of aggregate-level CSMs (ACSMs), and particularly the widespread use of median SGPs (Betebenner, 2010a), motivate a critical review of ACSM properties. We introduce a framework for “simple” choices among ACSMs, including the choice of aggregation function, the choice of quantile vs. OLS regression, and the choice of fixed versus random effects. To illustrate these contrasts clearly, we restrict our focus to relatively simple ACSMs that condition only on past scores. We begin with the median SGP and contrast it against simple alternatives: 1) the mean SGP, 2) aggregates of Percentile Ranks of Residuals (PRRs; Castellano & Ho, 2012)—the OLS-regression analog to SGPs—and 3) simple versions of “value-added” models—the Fixed-Effect and Random-Effect metrics.

Within this third class of models, our approach contrasts with those often taken in the large and growing body of literature on variability in value-added scores over time or across measures (e.g., McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2011) by focusing exclusively on the most elemental versions of these models that condition only on past scores. This defers the problem of “value added” inferences and allows focus on the empirical and theoretical impact of simple choices among metrics that support inferences specifically about aggregate-level conditional status. We describe the impact of these simple choices on three practical factors 1) school rankings, 2) robustness across varying the number of years of prior scores, and 3) scale transformation invariance. Together, these analyses represent an effort to better understand the theoretical and practical differences between commonly used ACSMs, as well as the practical variability of any single ACSM under plausible alternative specifications.

Aggregate-Level Conditional Status Metrics
An ACSM is a measure that describes the “current grade” scores of a group of students in terms of conditional expectations given “prior grade” scores. To follow a common reporting scenario in state testing programs, we set conditional expectations using only prior grade scores, but, strictly speaking, any student- or group-level variable could be used instead of or in addition to prior grade scores.

Two alternative classes of metrics provide a useful contrast to ACSMs. The first consists of gain- or slope-based models. These require data that share a common scale over time, allowing for meaningful gain scores in the form of \(X_2 - X_1\), where \(X_i\) represents a score at time \(i\). Longitudinal data with more than two time points permit the plotting of individual or average score trajectories over time. Linear or nonlinear regression models (e.g., growth curve models) fit to these data can support inferences about average and aggregate growth. Although both gain-based metrics and ACSMs may use regression models, gain-based metrics model scores over time, whereas ACSMs model current grade scores as a function of prior grade scores. ACSMs do not make any temporal distinction between past grades nor require a common scale over time.

A second, contrasting class of metrics is supported by so-called multivariate models. This class includes the cross-classified model of Raudenbush and Bryk (2002), the variable persistence model of McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004), and the Education Value-Added Assessment System (EVAAS) model of Sanders, Saxton, and Horn (1997). Whereas ACSMs describe current grade scores in conditional distributions of prior grade scores, multivariate models simultaneously consider the full trajectories of students over time. These multivariate models are considerably more complex and are used largely in support of “value added” inferences that attribute higher- or lower-than-expected aggregate student performance to causal effects of particular teachers and schools. In contrast, ACSMs are readily
interpretable without these value added, causal inferences and are relatively simple to estimate and describe, which may explain their widespread use in practice.

McCaffrey et al. (2004) usefully describe gain-based models and conditional status models as special cases of multivariate models, and we recommend their presentation as a broader statistical framework. A full discussion of “growth” and “value added” inferences is beyond the scope of this paper. We take the position that ACSMs can inform discussions of both growth and value added, but they should first be interpreted in terms of the conditional status interpretations that they support most directly.

**Distinguishing among Aggregate-Level Conditional Status Metrics**

Table 1 lists the seven ACSMs addressed in this paper: median Student Growth Percentiles (medSGPs), mean SGPs (meanSGPs), median Percentile Ranks of Residuals (medPRRs), mean PRRs (meanPRRs), Residuals from Mean Regression (RMRs), the Fixed-Effects Metric (FEM), and the Random-Effects Metric (REM). Table 1 also classifies these models by identifying principles along which these ACSMs differ, reflecting the “simple choices” implicit in ACSM selection.

The first identifying principle is the regression approach that establishes expectations for the current grade scores. The models supporting ACSMs use different regression approaches including simple Ordinary Least Squares (OLS) regression, quantile regression, and multilevel regression. The second identifying principle is the aggregation function, with some metrics supporting aggregate-level inferences through mean- or median-based operations. The third identifying principle is the order in which conditioning and aggregation occur. Table 1 refers to this as the “order of operations” and classifies ACSMs by one of three approaches: condition-then-aggregate, aggregate-then-condition, or simultaneous. The fourth identifying principle is the
scale for interpretation, where some metrics are expressed as percentile ranks and others are expressed on the score scale of the current grade test (i.e., as residuals).

Additional Conditional Status Metrics and Alternative Frameworks

Table 1 does not represent an exhaustive list of either metrics or distinguishing principles. Clearly, there are many other aggregation functions and possible scales for interpretation. Selection of the seven ACSMs is motivated by widespread use, in the case of medSGPs, FEMs, and REMs, and illustrative contrasts, in the case of meanSGPs, aggregated PRRs, and RMRs. Although there are bodies of literature supporting individual ACSMs, there are few examples of theoretical and empirical comparisons. McCaffrey et al. (2004) presented a larger framework for “value-added” models, but we focus on what they called “covariate adjustment” metrics and emphasize contrasts within this class.

At the individual level, Castellano and Ho (2012) compared SGPs and PRRs. They found that these metrics are highly similar across a range of simulated and real data scenarios. They established guidelines for sample size, the number of prior grades required for stable estimates, and the relative degree of scale-transformation invariance for the two metrics. This paper extends their work to the aggregate level and broadens the scope of metrics under consideration.

A key reference and the only one we know that compares properties of ACSMs is Briggs and Betebenner’s (2009) investigation of the scale invariance of medSGPs. They find that correlations between medSGPs derived from extant and transformed scales are near unity, a finding consistent with scale invariance. They also compare the scale invariance of medSGPs to that of aggregate-level effects from a “layered” value-added model, following Ballou, Sanders, and Wright (2004), that models individual score vectors as a joint multivariate outcome.
This paper differs from the Briggs and Betebenner (2009) paper in three ways. First, we expand their framework by considering medSGPs as one of many possible ACSMs. Second, we evaluate scale invariance with different transformations and use ranges instead of correlations as an outcome statistic. Third, we do not consider the layered model, as its construction as a multivariate model is fundamentally distinct, statistically and substantively, from the conditional status interpretations we focus on in this paper. We describe each ACSM of interest in the following subsections.

**Median and Mean Student Growth Percentiles**

The SGP metric uses quantile regression to describe students’ current status in the context of their past performance. The estimation of individual SGPs is described in detail by Castellano and Ho (2012) and is documented in the R (R Development Core Team, 2009) package “SGP” by Betebenner (2010b). We briefly review the procedure here and transition to properties of two aggregate-level SGPs: median and mean SGPs.

The quantile regression approach can be clarified by contrasts to simple OLS regression of current grade scores on prior grade scores. In OLS regression, this prediction takes the form of a conditional mean: an average current score for students with a particular prior grade score. In contrast, quantile regression allows for the estimation of conditional quantiles, such as the median, 25th percentile, and 90th percentile, for students with a particular prior grade score. Neither regression approach requires current and prior grade scores to be on the same scale.

The SGP estimation procedure involves estimation of 100 conditional quantile surfaces corresponding to quantiles from .005 to .995 in .01 increments (Betebenner, 2010b). These surfaces represent boundaries. Students with observed scores that fall between two adjacent surfaces are assigned an SGP represented by the midpoint quantile between these boundaries.
For example, a student whose current grade score is between the .495 and .505 predicted quantile surfaces has an SGP of 50.

Just as OLS regression can include nonlinear functions, the SGP quantile approach employs nonlinear B-spline functions to fit conditional quantile lines to accommodate nonlinearity and heteroscedasticity in the data (Betebenner, 2009). We follow this approach, implemented in Betebenner’s (2010b) “SGP” package. We also follow Castellano and Ho (2012) in increasing the resolution of SGPs by estimating 1000 instead of 100 lines, for quantiles from .0005 to .9995 by .001, allowing reporting of SGPs to one decimal point instead of as integers.

This paper contrasts two SGP-based ACSMs: medSGPs and meanSGPs. The medSGP and meanSGP metrics involve simple aggregation of SGPs using the median and mean functions, respectively. Because the aggregation follows after the application of the regression model, Table 1 classifies SGP-based ACSMs as metrics that condition first and then aggregate.

The SGP-based ACSM in operational use is the medSGP, following Betebenner’s (2008b) recommendation to use medians due to the ordinal nature of percentile ranks. Means, in contrast, are computed under the assumption that an interval scale underlies the averaged units. Such an assumption is violated by the percentile rank scale whenever the underlying, latent distribution for which percentile ranks are reported is non-uniform. However, we take the view that equal-interval scale properties should be evaluated with respect to uses and interpretations. The statistical features of means may support useful inferences and properties even when scales do not appear to have equal-interval properties (Lord, 1956; Scholten & Borsboom, 2009). Further, strict equal-interval properties are rare among all test score scales (Spencer, 1983; Zwick, 1992), and this has not stopped the common practice of averaging test scores. Threats to
interpretations of ordinal-based averages are a matter of degree and may be offset by the advantages of averages, as we will demonstrate with an analysis of sampling variability.

**Median and Mean Percentile Ranks of Residuals**

Castellano and Ho (2012) contrast SGPs with an analogous approach that uses the percentile ranks of residuals (PRRs) from an OLS regression of current scores on past scores. This approach has been used previously in a range of applications (e.g., Ellis, Abrams, & Wong, 1999; Suen, 1997; Fetler, 1991). A student’s PRR and SGP are both interpretable on a percentile rank scale as conditional status given past scores, but PRRs require less computation time and are anchored in an elementary statistical framework. Castellano and Ho (2012) show that SGPs and PRRs have correlations of about 0.99 for real data and root mean square differences of approximately 3 on the percentile rank scale. They also demonstrate that PRRs recover simulated “benchmark” percentile ranks better than SGPs when linear regression assumptions hold, and they identify skewness levels at which SGP recovery exceeds PRR recovery.

Like SGPs, PRRs are student-level statistics, but, following the condition-then-aggregate approach, they are easily aggregated to meanPRRs and medPRRs. We adopt Castellano and Ho’s (2012) notation for PRRs to emphasize that simple regression models are models for conditional means, in this case, the mean of a current grade score variable, $Y$, conditional on scores from $J$ prior grades, $X = X_1, X_2, ..., X_J$:

$$\mu_{Y|X} = \alpha + \sum_{j=1}^{J} \beta_j^{(T)} X_j.$$  \hspace{1cm} (1)

Here, $\alpha$ is the intercept, and the $\beta_j^{(T)}$ parameters are the regression coefficients for each prior test score included. The superscript $(T)$ distinguishes these regression coefficients as arising from a “total” regression using all individual (disaggregated) data, a contrast with upcoming models. As
a model for the conditional mean, this approach contrasts with the quantile regression model underlying SGPs. The SGP quantile regressions are nonlinear, and there are typically 100 quantile regression functions (and in our case, 1000) instead of one OLS regression function.

The residuals for the regression shown in Equation 1 are the simple differences between observed values and the expected values given past scores: $Y - \hat{\mu}_{Y|X}$. The PRRs are the residuals rank ordered and transformed to percentile ranks, which we round to one decimal point for consistency with the granularity of SGPs. Although studentized residuals could be used, we follow Castellano and Ho (2012) and use raw residuals for simplicity and because differences are inconsequential at the aggregate level. We aggregate PRRs over students within groups using the mean and median function to obtain meanPRRs and medPRRs respectively. This paper investigates aggregate-level PRRs to support better understanding of the metric in its own right and provide an illustrative contrast to the more widely used medSGP.

Residuals from Mean Regression

The SGP and PRR metrics condition and then aggregate in two distinct steps. The RMR metric fits a regression model to aggregate values, aggregating first, then conditioning. The RMR metric thus represents the conditional status of aggregates instead of an aggregate of conditional status. As the name suggests, RMRs are the residuals from the regression of average current grade scores on their $J$ prior grade average scores. This regression model is often referred to as a “between-groups” regression, as it ignores any within-group variability (e.g., Snijders & Bosker, 2011). RMRs are found using the same OLS regression as for PRRs, but the individual prior grade scores are replaced by their group averages, reducing the fitted data from $n$ students to $G$ groups. This is a simple substitution of mean scores ($\bar{X}$) for individual scores in Equation 1:
\[ \mu_{\bar{y}|\bar{x}} = \alpha + \sum_{j=1}^{J} \beta_j^{(B)} \bar{X}_j. \]  

Here, \( \bar{X}_j \) represents the mean scores for the \( j^{th} \) prior grade, \( \beta_j^{(B)} \) refers to the corresponding regression coefficients from the “between-groups” regression, and \( \alpha \) denotes the intercept. The RMR is the difference between an observed current grade average score (\( \bar{y}_g \)) and its expected current grade average score (\( \bar{\mu}_{y|\bar{x}} \)).

**Fixed-Effects Metric**

The model supporting FEMs is sometimes described as a “covariate adjustment” model or, more simply, as an Analysis of Covariance (ANCOVA). The separation of aggregation and conditioning into two distinct steps breaks down in these “multilevel” models, and the order of aggregation and conditioning is better described as simultaneous. In this approach, an aggregate-level “fixed effect” is added to the individual-level OLS regression of current grade scores on \( J \) prior grade scores. The fixed effects can be represented by the coefficients (\( \gamma_g \)) of dummy variables for groups, resulting in distinct intercepts for each group \( g \):

\[ \mu_{y|x} = \gamma_g + \sum_{j=1}^{J} \beta_j^{(W)} X_j. \]  

This model constrains all groups to have the same slope estimates (\( \hat{\beta}_j^{(W)} \)) estimated from what is often called a “within-groups” regression (e.g., Snijders & Bosker, 2011). Standard errors and other statistical properties of FEMs are easily calculated in an OLS regression framework.

FEMs contrast usefully with the other ACSMs based on OLS regression: aggregated PRRs and RMRs. The underlying models are the within-groups, total, and between-groups regressions, respectively. The multilevel modeling literature provides a useful conceptual link between the three. When group sizes are equal and there is one prior grade predictor (\( X \), the total
regression coefficient, $\beta^{(T)}$, is the weighted sum of the between- ($\beta^{(B)}$) and within-group ($\beta^{(W)}$) regression coefficients, $\beta^{(T)} = \eta^2 X \beta^{(B)} + (1 - \eta^2 X) \beta^{(W)}$, where $\eta^2 X$ is the ratio of the intraclass correlation coefficient for $X$ to the reliability of its group mean (Snijders & Bosker, 2011, p. 31).

The focus of this paper is not on the regression coefficients but the contrasts between the ACSMs themselves. There is a great deal of empirical and conceptual similarity between FEMs and simple averages of OLS residuals (without the percentile rank transformation). The FEMs themselves can be usefully described as the result of an analogous three-step procedure, where current and prior grade scores are centered on group averages, the centered current grade scores are regressed on the centered prior grade scores, and residuals (computed using the uncentered prior grade scores) are averaged within groups. Accordingly, the mean is listed as the effective aggregation function for FEMs in Table 1. This also illustrates how the FEM ordering of conditioning and aggregation is not straightforward and thus classified as simultaneous.

**Random-Effects Metric**

The model supporting the REM is a “random intercepts” model or a random-effects ANCOVA (Raudenbush & Bryk, 2002). It is similar to the model supporting FEMs, but the estimation procedure assumes that the group-level effects are uncorrelated with the student prior-grade scores. Although this assumption is unlikely to hold in practice, we include REMs for their illustrative contrasts to the other metrics and because random-effects models are often used in value-added modeling despite violations of this assumption (Kim & Fees, 2006). Given school-level random intercepts designated as $u_g$ and an overall average group intercept designated as $\gamma_0$:

$$\mu_{Y|X} = (\gamma_0 + u_g) + \sum_{j=1}^{J} \beta^{(R)}_j X_j. \quad (4)$$
This random intercepts model constrains the slopes \( \hat{\beta}_j^{(R)} \) to be equal across groups, but the estimated slopes will differ from those of the FEM model \( \hat{\beta}_j^{(W)} \). As previously explained, \( \hat{\beta}_j^{(W)} \) can be estimated with a simple regression of current on past grade scores following group-centering. Wooldridge (2010) draws a useful contrast by demonstrating that \( \hat{\beta}_j^{(R)} \) can be estimated by partial group-centering, resulting in the estimates of \( \hat{\beta}_j^{(R)} \) and \( \hat{\beta}_j^{(W)} \) being most similar when group sizes are large or when the proportion of between-group variation is small.

The REM metric differs from the FEM not only due to the differing estimates of \( \beta_j \) but also due to a “shrinkage factor,” where REMs are shrunk to the overall mean. For any given dataset, there will be more shrinkage for smaller groups than larger groups. Across datasets, there will be more shrinkage for data with more within-group than between-group variation.

**Visual Comparison of ACSMs**

Figure 1 illustrates key distinctions between ACSMs through visualizations of the metrics on a scatterplot of real data. The figure depicts a scatterplot with a cohort’s Grade 6 scores plotted on their Grade 5 scores. Light gray conditional boxplots show the empirical bivariate distribution of current year scores given each score point in the previous year. The OLS regression line in dark gray has a slope \( \hat{\beta}_1^{(T)} \) (Equation 1), and it serves as a reference for aggregate PRRs that convert residuals into percentile ranks. Points above the line represent students who have high PRRs, or performed better in Grade 6 than expected given their Grade 5 scores. Accordingly, groups with more students above the line will receive higher ACSMs than groups with more students below the line.

The median quantile spline in black serves as a reference for aggregate SGPs. There are 1000 of these quantile splines that map individual data points onto the percentile-rank-based
SGP scale. Points closest to this line will have SGPs of 50. Like PRRs, points above the line will be assigned higher SGPs. The jagged shape of the spline at the extreme arises from corrections that ensure that none of the quantile regression lines will cross each other (Betebenner, 2010b; Castellano & Ho, 2012; Dette & Volgushev, 2008).

The dashed line denotes the regression line for the RMRs with slope $\hat{\beta}_1^{(B)}$ (Equation 2). The students in this dataset belong to 546 schools, and the solid, gray squares represent their school-level average Grade 6 scores plotted against their school-level average Grade 5 scores. The slope of the RMR line is steep compared to the reference lines of the other models in part because it is fit to school-level scores that have a higher correlation than the individual student data. Solid, gray squares above this line represent schools with positive RMRs, whose average Grade 6 scores are higher than expected given their average Grade 5 scores.

The fixed effects model supporting FEMs fits different regression lines to each school’s individual-level data while constraining slopes $\hat{\beta}_1^{(W)}$ to be equal (Equation 3). One mid-ranking school can be selected as a reference school; this school’s regression line is shown as a dotted line. All other schools have parallel regression lines (not shown), and vertical distances between the regression line of any school and that of the reference school is that school’s FEM.

A similar line for the REM model could have been plotted in Figure 1, but it would be visually indistinguishable from the plotted FEM line. The REM model produces an overall regression function with slope $\hat{\beta}_1^{(R)}$ and allows prediction of random effects $\tilde{u}_g$ for each school (Equation 4) that can be added to the overall intercept to produce school-specific parallel lines.

The PRR, FEM, and RMR regression lines correspond to total, within-group, and between-group regressions respectively. As the previous section described, the total regression coefficient supporting PRRs, $\hat{\beta}_j^{(T)}$, will be closer to the within-group coefficient supporting
FEMs, $\hat{\beta}_j^{(W)}$, when between-group variation is low. This is clearly seen in Figure 1, where the slope of the regression line for PRRs is closer to that of FEMs than that of RMRs. Despite possible similarities between the $\hat{\beta}_j^{(T)}$ and $\hat{\beta}_j^{(W)}$ coefficients, the PRR and FEM metrics will differ for reasons shown in Table 1 and at magnitudes illustrated later in this paper.

**Theoretical Differences between Mean and Median SGPs and PRRs**

The aggregated SGP and PRR metrics all share the same order of operations—condition-then-aggregate—and scale for interpretation—the percentile rank scale—making them a convenient subclass of metrics to identify contrasts within. Before reporting empirical similarities and differences between each pair of ACSMs, we first describe the sampling distributions of the popular medSGP and its close relatives, the meanSGP, medPRR, and meanPRR, to inform the choice between mean- and median-aggregated ACSMs.

As percentile ranks, SGPs and PRRs follow a theoretical uniform distribution with a 0 to 100 range and a standard deviation of $50/\sqrt{3}$. The sampling distribution of medians (as the $(n + 1)/2$ order statistic) from a uniform distribution is known to follow a beta distribution, under random assignment of students to groups of size $n$ (Casella & Berger, 2001). The expected value of medSGPs and medPRRs under these conditions is unsurprisingly 50, and their standard error is $50/\sqrt{n + 2}$. In contrast, under random assignment of students to groups, the standard error of the meanSGP and meanPRR is simply $50/\sqrt{3n}$. Accordingly, the medSGP and medPRR have a theoretical standard error that is $\sqrt{3n/(n + 2)}$ times larger than the standard error of the meanSGP and meanPRR. This factor approaches $\sqrt{3} \approx 1.73$ for large $n$.

The ratio of variances is known as the asymptotic relative efficiency (ARE) and affords statements like, mean SGPs and PRRs are 3 times more efficient than their median counterparts. Equivalently, the sample median, in this case, requires 3 times as many observations to be as
efficient as the sample mean. Even for \( n = 20 \), the minimum subgroup size for reporting in
many states (e.g., Colorado Department of Education [CDE], 2012; Massachusetts Department
of Elementary and Secondary Education [MDESE], 2009), the theoretical standard error of a
randomly sampled medSGP (or medPRR) and meanSGP (or meanPRR) are 10.66 and 6.45
respectively on the percentile rank scale, and the ARE of means over medians is 2.73.

We conducted simulations to address two practical extensions of these findings. First,
because SGPs are reported as integers instead of as continuous variables, we investigate ARE for
discrete integers from 1 to 99. Under random sampling and for large \( n \), the ARE of means over
medians is closer to 4. Second, to address nonrandom assignment, we computed bootstrapped
standard errors for meanPRRs and medPRRs under varying intraclass correlations (ICCs)
ranging from 0 to .4—values expected (or slightly larger than expected) of educational data
(Hedges & Hedberg, 2007).

For homogeneous group sizes of \( n = 20, 50, \) and 100 and across 1000 bootstrap samples
under each condition, we found that the ratios of error variances of medPRRs to meanPRRs
ranged from about 2.30 to 3.10, with the smaller relative efficiencies corresponding to the larger
ICCs. That is, at best, when 40 percent of the total variation of the current score is attributed to
between-group variation, median-SGPs or PRRs require 2.30 as many observations to match the
efficiency of their mean-based counterparts. The magnitude of these differences in sampling
variability represents a strong argument for using mean-based SGPs or PRRs, even on an
unequal-interval scale, to increase the stability of estimates of aggregate-level conditional status.

**Data**

We use the same two statewide data files as those from the real-data analyses of
Castellano and Ho (2012), allowing for a common data reference point between these papers.
There are a total of four distinct four-year complete-record longitudinal datasets: two states crossed with two subjects. The states, referred to as “State A” and “State B,” contrast usefully in size and scaling procedures. The State A dataset contains complete records for a single cohort of about 25,000 students with reading and mathematics scores from grade 3 (2004-2005) to grade 6 (2007-2008) on a vertical scale with increasing variance over time. The State B dataset has mathematics and reading scores for a cohort of about 75,000 students from grade 3 (2002-2003) to grade 6 (2005-2006). State B does not have a vertical scale.

The State A dataset allows for district-level analyses, whereas the State B dataset allows for school-level analyses. Both datasets include four years of data from a single cohort representing sixth graders in the “current” year. For convenience, we describe these groups as “schools” instead of district-level grade cohorts and school-level grade cohorts respectively.

To adhere to standard reporting rules, we follow Colorado’s cutoff for reporting medSGPs (CDE, 2012) and exclude results for schools with fewer than 20 students. For aggregate-level SGPs and PRRs, we follow Colorado’s practice of including all students in individual-level computations but excluding students from small schools from aggregate-level analyses. In contrast, RMRs, FEMs, and REMs use aggregate-level information early in estimation; thus, for these metrics, we exclude students from small schools prior to any calculations. Both of these practices are consistent with a policy of implementing exclusions just prior to aggregation.

This decision rule excludes about 18 percent of districts and 2.8 percent of students in State A and about 20 percent of schools and 1 percent of students in State B. These levels of exclusion do not threaten the overall conclusions drawn from the cross-metric comparisons shown here. In State A, the median and mean district-cohort sizes are 46 and 90 respectively,
with a maximum size of 1700. In State B, the median and mean school-cohort sizes are 133 and 139 respectively, with a maximum size of 380.

### Cross-Metric Comparability

The previous sections contrasted ACSMs by their statistical models and the identifying principles in Table 1. This section describes the magnitude of the practical impact of ACSM selection in terms of three statistics. First, root mean square deviations (RMSDs) describe the expected absolute difference across metrics that share a common scale. Aggregate-level SGPs and PRRs share a common percentile rank scale, thus Table 2 displays pairwise RMSDs for these metrics. For example, the RMSD comparing meanPRR and meanSGP metrics follows:

\[
\text{RMSD}_{\text{meanPRR}, \text{meanSGP}} = \sqrt{\frac{\sum_{g=1}^{G}(\text{meanPRR}_g - \text{meanSGP}_g)^2}{G}}.
\]  

(5)

Second, Table 3 displays Spearman Rank correlations for all school-level metrics. Third, Figure 2 displays the expected absolute difference in percentile ranks associated with choices between two chosen reference metrics. This section discusses each of these in turn. Following the practice of many states (e.g., Sanders, 2006), this analysis is restricted to ACSMs calculated with \( J = 3 \) prior year scores.

### Comparing Aggregate-Level SGPs and PRRs

These first four metrics from Table 1 contrast in terms of regression model (OLS vs. quantile) and aggregation function (median vs. mean). At the individual-level, Castellano and Ho (2012) show strong linear associations and small absolute differences between SGPs and PRRs. Using the same State A and State B datasets, using \( J = 3 \) prior year scores, they show correlations around 0.99 and RMSDs between 2 and 5 percentile ranks.
At the aggregate level, Table 2 can be summarized in three findings. First, the meanPRR and meanSGP metrics have small absolute differences, with RMSDs between 0.40 and 0.61. Second, the medPRR and medSGP metrics are relatively more discrepant, with RMSDs between 1.33 and 2.08. Third, pairwise comparisons of mean-based metrics and median-based metrics reveal the most dissimilarity at about 4 percentile ranks, even when two metrics share a regression function. These results suggest that the relative impact of the choice of aggregation function is far greater than the relative impact of the choice of regression model. The results are also consistent with the previous findings of the ARE of means over medians, where mean-based metrics are more stable across regression functions than median-based metrics. Table 3 presents Spearman Rank correlations that lead to consistent conclusions. Mean-mean comparisons are most similar, followed by median-median comparisons, followed by median-mean comparisons regardless of the regression model.

Comparing All ACSMs

The lower right quadrants of Table 3 show the practical impact of “simple” choices among ACSMs derived from multilevel models in terms of Spearman rank correlations. FEMs and REMs are the most highly correlated (.993 ≤ r ≤ .999). As noted earlier, these two metrics can disagree due to differing slope estimates and differential shrinkage due to school size. The large size of many schools and the small degree of between-group variation lead to similar slopes, $\hat{\beta}_j^{(R)}$ and $\hat{\beta}_j^{(W)}$, for the models underlying the FEMs and REMs. Additionally, the differential shrinkage due to varying school sizes does not result in substantial changes in rankings. The FEMs/REMs are not as strongly correlated with RMRs (.937 ≤ r ≤ .973) due to the small amount of between-group variation that leads to the differing slopes (see Figure 1).
The upper right and lower left quadrants of Table 2 show comparisons between SGP/PRR-based metrics and the FEMs, REMs, and RMRs. The lowest correlations are between the RMR metric and median-based SGPs/PRRs. Also, the meanSGPs/meanPRRs generally have relatively lower correlations with the RMRs than the other metrics. The relatively weak relationships between RMRs and other ACSMs suggest that ignoring all individual-level information results in empirically distinct conditional status interpretations. The conditional status of averages is distinct from average conditional status, both in theory and in practice.

The Table 2 correlations also show that the FEM and REM metrics have higher associations with mean-based SGPs and PRRs ($r \approx .99$) than median-based SGPs and PRRs ($0.90 \leq r \leq .98$). If aggregate-level SGPs and PRRs are intended to be ad hoc approximations to the more comprehensive statistical framework of multilevel models, it is clear that mean-based metrics result in closer approximations than median-based metrics.

### Absolute Differences in School Percentile Rank

Correlations provide a limited perspective on differences among metrics, as coefficients are generally high but difficult to interpret on a practical scale. Alternatively, we can compare two metrics using absolute differences in the percentile ranks of schools. If a school drops from the 99th percentile on one metric to the 1st percentile on another, the absolute difference in percentile ranks is 98. Large absolute differences indicate dissimilarity between metrics on a policy-relevant scale. These percentile ranks should not be confused with aggregate-level SGPs and PRRs that are also on the percentile rank scale. As an example, the medSGP of one school is 32, which is greater than or equal to the medSGPs of 36% of all the schools, so the school is at the 36th percentile on the medSGP metric. The same school has a meanSGP of 40, which
corresponds to the 26th percentile on the meanSGP metric. The absolute difference in percentile ranks for this school is thus $|36 - 26| = 10$.

Figure 2 uses boxplots to show the distribution of differences between ACSMs and two “reference metrics,” medSGPs and FEMs. We select these metrics as reference metrics because the medSGP metric is widely used in practice and the FEM metric is widely used in the research literature. The former metric is more descriptive, and the latter metric exists in a statistical framework that explicitly incorporates aggregation. Boxplots in each panel are ordered in descending order of similarity to their reference metric, as measured by the magnitude of the correlations in Table 3. This plot uses the 546 schools from the State B Reading data, but any of the other datasets could have been used to illustrate the same findings.

The number to the left of each boxplot refers to the median absolute difference in school percentile ranks. This can be loosely interpreted as the typical difference that switching between metrics would make in terms of percentile ranks. For example, switching between medSGPs and meanSGPs would change a typical school’s ranking by 4 percentile ranks. The boxplots also show that, at worst in this dataset, switching between medSGPs and meanSGPs changes a school’s ranking about 30 percentile ranks. Figure 2 supports similar interpretations across all pairwise comparisons. The typical change in rank may not be large, but, as indicated by the whiskers of the boxplots, a small number of schools will change by 20 percentile ranks or more.

Figure 2 is consistent with Tables 2 and 3 in dividing the metrics into four categories by similarity. First, the REM and FEM metrics are highly similar, as are the meanSGP and meanPRR metrics. Second, these two pairs of metrics are very similar to each other, where the right panel of Figure 2 shows that both meanSGPs and meanPRRs have a median difference in percentile ranks of 2 from FEMs. The left panel of Figure 2 shows that the third category of
metrics is the relationship between medSGPs and medPRRs, as all other metrics differ more dramatically from these metrics. Finally, the distinctive RMRs are in a category of their own.

An alternative approach to describing metric dissimilarity is to describe the proportion that stay in the same quartile or decile, or count the number of schools that change one or more quartiles. These metrics are intuitive and practical but depend on the number and location of cut scores. For example, a school with a cross-metric change of 20 percentile ranks may or may not cross a quartile boundary. Figure 2 is a more robust representation that describes the magnitudes of typical and extreme changes in school percentile ranks.

**Sensitivity to the Number of Prior Grades of Test Score Data**

A consequential characteristic of CSMs is that the selection of predictors will affect conditional distributions. As conditional distributions change, so will metrics that reference them. In this context, school ACSMs will depend on which and how many prior grades, or, equivalently, “prior years” (Castellano and Ho, 2012), data users choose to include as predictors. This section quantifies this “prior-grade dependence” and compares them across metrics.

We compute the 7 ACSMs of interest using \( J = 1, 2, \) and 3 prior grades, holding the Grade 6 scores fixed as the current score or outcome variable, \( Y \) for our four statewide datasets. Accordingly, the \( J = 1 \) dataset conditions on Grade 5 scores, the \( J = 2 \) dataset on Grade 4 and 5 scores, and the \( J = 3 \) dataset on scores from Grades 3 to 5. Castellano and Ho (2012), at the individual-level, found that that SGPs and PRRs show similar degrees of prior-year dependence. This dependence is smallest when comparing \( J = 2 \) to \( J = 3 \) prior years of data (RMSDs between 5 and 7) and largest when comparing \( J = 1 \) versus \( J = 3 \) prior years of data (RMSDs between 11 and 15). We replicate this analysis at the aggregate level and extend it to other ACSMs.
Figures 3(a) and (b) illustrate the RMSDs for aggregated SGPs and PRRs across different prior grade specifications for State A and State B Reading datasets respectively. The plots for the Math datasets look almost identical to those for Reading and are not included here. These plots reveal a similar pattern to the individual-level results reported by Castellano and Ho (2012) but, expectedly, with much smaller RMSDs due to aggregation. The groups of bars are ordered from smallest to largest with smallest prior-grade dependence again found for $J = 2$ versus $J = 3$. More generally, prior-grade dependence is smallest when comparing two models with small differences in the number of prior-grades (small $\Delta J$) and when there are many prior grades of data (large $J$). This is not a surprising finding in the context of multiple regression when predictors have similar and positive pairwise correlations with each other and the outcome.

Figure 3 also illustrates the effect of prior-grade dependence by type of aggregated SGPs and PRRs. There is little difference between PRR- and SGP-based ACSMs on the criterion of prior-grade dependence, that is, the bars for meanSGPs and meanPRRs are nearly the same height and likewise for medSGPs and medPRRs. But Figure 3 clearly depicts greater prior-grade dependence for median-based SGPs and PRRs than mean-based SGPs and PRRs, by a factor of about 2 on the RMSD scale. If consistency across prior-grade specifications is a high priority, meanSGPs and meanPRRs will be preferable over their median-based counterparts.

Figure 3 does not show results from the FEM, REM, and RMR metrics, as these are reported on a different scale that does not support meaningful RMSDs. However, in terms of correlations, these three metrics have similar degrees of prior-grade dependence, and the magnitudes of prior-grade dependence follow the same general pattern as that seen in Figure 3. Correlations are largest between the $J = 2$ and $J = 3$ prior grades, ranging from about .97 to .99,
and smallest between $J = 1$ and $J = 3$ prior grades, ranging from about .86 to .97. We exclude this figure to avoid redundancy.

**Scale Invariance**

A desirable property of academic performance measures is robustness to plausible monotonic scale transformations of the test score data. As the construction of a test scale involves a degree of subjectivity, and because different test scales can be related by monotonic scale transformations, we can quantify the scale-dependence of target metrics through plausible transformations. One of the motivations behind quantile-regression-based SGPs is invariance under scale transformations (Betebenner, 2009). Quantile regression is theoretically scale-invariant to monotonic transformations of at least the dependent variable (Koenker, 2005). In contrast, OLS regression is not invariant to nonlinear transformations of variables, particularly if they distort the linear relationships among the variables and diminish model fit.

Castellano and Ho (2012) found that for both simulated multivariate normal test score data and empirical data, $J=1$ prior-grade SGPs were markedly more invariant to $J=1$ prior-grade PRRs under a set of monotonic scale transformations. They reported that SGPs tended to vary within 1 or 2 percentile points of each other across the transformations, whereas students’ PRRs varied within about 4 to 5 percentile points on average.

Briggs and Betebenner (2009) compared the scale invariance of medSGPs and aggregate-level effects from a “layered” value-added model. They used transformations that reflected linear or nonlinear growth, constant or increasing variance, and the more extreme exponential transformation. They found that for $J = 1$, 2, and 3 prior grades, medSGPs were near perfectly correlated across the transformations, whereas VAM effects were less strongly correlated ($r > .9$) with smaller correlations arising from the exponential transformation ($r \approx .3$ to .6). Our scale
invariance analysis differs from Briggs and Betebenner’s (2009) in the choice of scale transformations, the inclusion of aggregated PRRs, meanSGPs, RMRs, FEMs, and REMs, the exclusion of the non-ACSM multivariate model, the use of multiple empirical datasets, and the choice of outcome statistics. We avoid correlations because they are limited to pairwise expressions of differences, and we implement more than two transformations. Thus, we quantify the amount of transformation-induced variability by using mean ranges on a common scale of standard deviation (SD) units.

Our transformations are piecewise transformations of the individual grade-level scores that stretch or shrink parts of the scale, resulting in four sets of transformations that result in four transformed datasets: “positive skew,” “negative skew,” “positive kurtosis,” and “negative kurtosis.” We use the same skewness, \( S(z) \), and kurtosis, \( K(z) \), transformations defined in Castellano and Ho (2012) and reproduced below:

\[
S(z) = \begin{cases} 
(z/k) + (1 - k)/(k \times (k + 1)); & z < -1 \\
(2 \times z)/(k + 1); & -1 \leq z < 0 \\
(k + 1) \times z]/2; & 0 \leq z < 1 \\
(k \times z) + (1 - k)/2; & z \geq 1 
\end{cases}
\]

\[
K(z) = \begin{cases} 
(z \times k) + k - (1/k); & z < -1 \\
z/k; & -1 \leq z < 0 \\
z/k; & 0 \leq z < 1 \\
(z \times k) + (1/k) - k; & z \geq 1 
\end{cases}
\]

The constant \( k \) in these equations modulates the degree to which the skewness and kurtosis are altered. Following Castellano and Ho (2012), we use \( k = 1.2 \) for the positive skewness and kurtosis transformations and \( k = 1/1.2 \) for the negative skewness and kurtosis transformations. This mimics realistic values of skewness and kurtosis observed in practice, and these piecewise transformations follow those sometimes used by state testing programs (e.g. MDESE, 1999). The variable \( z \) represents the standardized grade-level test scores, where the scores are standardized by first subtracting the respective grade-level mean and then dividing by
the pooled SD across all the grade-levels of interest. This preserves the relative grade-to-grade variability in the empirical datasets. A final linear transformation is applied to the Grade 6 data: dividing by the Grade 6 SD. This has no effect on PRR- and SGP-based metrics and no statistical effect on the FEM, REM, or RMR metrics, but this allows for these latter metrics to be interpreted on a scale of individual-level SD units. We use $J = 3$ prior years for this analysis to reflect a common number of prior years used in practice.

The 7 ACSMs are estimated for all transformed datasets. For a given dataset, each school has five values for each ACSM—one each from the original, positive skew, negative skew, positive kurtosis, and negative kurtosis data. The FEM, REM, and RMR metrics are all interpretable in terms of individual-level SD units. To put the percentile rank-based metrics on the same SD unit scale, we simply divide by the known SD of a uniform distribution from 0 to 100, that is $\frac{50}{\sqrt{3}}$. Each school now has five values for each ACSM, and the range of these values is interpretable in individual-level SD units.

Figure 4 displays the average range of transformed ACSMs for schools in each empirical dataset. The most scale-invariant metric is the meanSGP metric, where an average school’s meanSGP range is less than .01 SD units under the performed transformations. This corresponds to less than one percentile rank point on the SGP scale. The REM, RMR, FEM, and meanPRR metrics are slightly more variable, or less scale invariant. The median-based metrics are noticeably more variable under the transformations. Consistent with previous analyses, the aggregation function is a more significant factor than the regression function in comparing scale invariance by ACSMs. The ordering is fairly consistent across the datasets, although State A has larger mean ranges than State B. This is also consistent across analyses, where State A generally shows greater dissimilarities across metrics, greater prior-grade dependence, and greater
transformation-induced variability. These results are partly due to fewer schools, slightly higher
between-grade correlations, and a data structure with increasing variance across grades.

It may seem counterintuitive that mean-based metrics are more robust to transformations
than median-based metrics. Medians are more stable under monotonic transformations that skew
a score scale (as long as the median value is not itself transformed). In this analysis of PRRs and
SGPs, the transformations are applied to the test score scale, not to the percentile rank scale of
SGPs and PRRs, which will and should remain roughly uniformly distributed between 0 and 100.
As transformations change individual SGPs and PRRs, the mean leads to more stability across
these changes than the median. Although it may seem contradictory, at the aggregate level, the
scale-dependent mean will maximize the scale-independent tendencies of SGPs.

Concluding Remarks

Although the literature on growth and value-added models is rich, focused empirical
comparisons on a clearly defined subset of models are rare. As we have argued, empirical
differences among gain-based models, multivariate models, and models supporting ACSMs are
expected, as each addresses different questions. We restricted our focus to ACSMs because their
models are similarly constructed to describe the status of a group by referencing current
performance to expectations given past scores. Empirical differences can thus be explained in
terms of the approaches that different ACSMs take to support conditional status interpretations.

Our real datasets support four empirically distinguishable categories of ACSMs. These
categories are (1) the group-specific intercepts (FEMs and REMs), (2) the mean-aggregated
SGPs and PRRs (meanSGPs and meanPRRs), (3) the median-aggregated SGP and PRRs
(medSGPs and medPRRs), and (4) the aggregate-then-condition metric (RMRs). The categories
are in order of decreasing similarity to the first category of group-specific intercepts. RMRs are
in a category by themselves given their distinct results from the other metrics for our particular
datasets. Although RMRs may be less distinct from the other ACSMs in datasets with greater amounts of between-group variation, the proportion of between-group variation in our datasets is not substantially different from those of other educational achievement datasets (Hedges & Hedberg, 2007). As an alternative to RMRs, we also evaluated a weighted-mean regression that weights school observations by school size, but this modification did not significantly improve associations between RMRs and other ACSMs.

An important caveat to this study’s empirical findings is that the relationship between SGPs and other ACSMs will decline as assumptions of linear regression models are violated. If the relationships between the current and prior grade scores exhibit extreme heteroscedasticity or nonlinearity, SGPs will provide more accurate conditional status inferences. Although this paper demonstrates robustness of findings across different states with different scaling procedures, a useful extension involves identification of pivot points where aggregated SGP metrics become markedly and systematically different from other ACSMs like Castellano and Ho (2012) did at the individual level. As they note, however, such studies do not universally promote one metric over others as much as emphasize the importance of selecting a model that fits the data.

A consistent finding is that the aggregation function is a more consequential decision than the regression function, and that there are marked differences between mean- and median-SGPs and PRRs. Although medians support interpretations of “typical values” and are theoretically more appropriate for a non-interval scale, this paper presents four findings that suggest considerable advantages of means over medians. The first is the theoretical asymptotic relative efficiency of the mean over the median under random assignment. The second is the observed empirical similarities between meanSGPs/meanPRRs and the FEM and REM metrics (Table 3), which follow from multilevel models that explicitly model aggregate conditional
status. Third, mean-based SGPs and PRRs are considerably more consistent as the number of prior-grade score variables varies (Figure 3). Lastly, mean-based SGPs and PRRs are more robust to scale transformations than their median-based counterparts (Figure 4).

In light of these findings, the widespread use of the medSGP statistic should be reconsidered. If concerns about the ordinal scale of percentile ranks remain, a simple alternative, that we denote as meanSGP, can be used that assumes that a normal distribution underlies the percentile rank function: meanSGP = \( \Phi \left( \sum \Phi^{-1} \left( \frac{SGP}{100} \right) / n \right) \), where \( \Phi \) is the standard normal cumulative distribution function and \( n \) is the number of students within a group. This aggregation function takes advantage of the benefits of means over medians while diminishing concerns about taking averages on rank scales. We implemented this alternative with SGPs and PRRs, but it is so highly correlated with simple meanSGPs and meanPRRs that it made no difference to our substantive conclusions, so we have adhered to the simpler specification.

Descriptions of ACSMs often incorporate the terms “growth” and “value added.” These terms are increasingly ambiguous, and we recommend reviews such as those by Castellano and Ho (in press) and Reardon and Raudenbush (2009) that specify assumptions and articulate the necessary data and model features to address specific questions about growth or value-added. We stress the importance of describing ACSMs in terms of what they literally do: summarize the performance of a group by referencing current status to expectations given past scores. Under this definition, dependencies are better anticipated. By understanding aggregate-level conditional status, it is less surprising that expectations for current performance should shift depending on the number of prior grades included in the model, or that the aggregation function is consequential. As stakes rise for accountability and evaluation decisions for teachers, schools, subgroups, and districts, clear descriptions of ACSMs and their dependencies become essential.
References


http://www.doe.mass.edu/mcas/tech/98techrpt.pdf


Table 1

Table of Aggregate-Level Conditional Status Metrics

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Metric</th>
<th>Regression Model</th>
<th>Aggregation Function</th>
<th>Order of Operations</th>
<th>Scale for Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>medSGP</td>
<td>Median Student Growth Percentile</td>
<td>Quantile Median</td>
<td>Condition-then Aggregate</td>
<td>Percentile Rank</td>
<td></td>
</tr>
<tr>
<td>meanSGP</td>
<td>Mean Student Growth Percentile</td>
<td>Quantile Mean</td>
<td>Condition-then Aggregate</td>
<td>Percentile Rank</td>
<td></td>
</tr>
<tr>
<td>medPRR</td>
<td>Median Percentile Rank of a Residual</td>
<td>OLS Median</td>
<td>Condition-then Aggregate</td>
<td>Percentile Rank</td>
<td></td>
</tr>
<tr>
<td>meanPRR</td>
<td>Mean Percentile Rank of a Residual</td>
<td>OLS Mean</td>
<td>Condition-then Aggregate</td>
<td>Percentile Rank</td>
<td></td>
</tr>
<tr>
<td>RMR</td>
<td>Residuals from Mean Regression</td>
<td>OLS Mean</td>
<td>Aggregate-then-Condition</td>
<td>Current-Year Score Scale (Residual)</td>
<td></td>
</tr>
<tr>
<td>FEM</td>
<td>Fixed Effects Metric OLS, ANCOVA/Multilevel</td>
<td>Mean Simultaneous</td>
<td></td>
<td>Current-Year Score Scale (Residual)</td>
<td></td>
</tr>
<tr>
<td>REM</td>
<td>Random Effects Metric Multilevel</td>
<td>Mean Simultaneous</td>
<td></td>
<td>Current-Year Score Scale (Residual)</td>
<td></td>
</tr>
</tbody>
</table>

Note: In all regression models, current status, or the current grade-level score, is the response variable and prior grade-level scores are the predictors.
Table 2

Root Mean Square Differences Comparing Each Pair of Aggregated Student Growth Percentiles and Percentile Ranks of Residuals

<table>
<thead>
<tr>
<th>Metric</th>
<th>meanSGP</th>
<th>medSGP</th>
<th>meanPRR</th>
<th>medPRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>State A Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meanSGP</td>
<td>4.24</td>
<td><strong>0.45</strong></td>
<td>4.44</td>
<td></td>
</tr>
<tr>
<td>medSGP</td>
<td>4.64</td>
<td>4.26</td>
<td><strong>2.08</strong></td>
<td></td>
</tr>
<tr>
<td>meanPRR</td>
<td><strong>0.40</strong></td>
<td>4.68</td>
<td>4.37</td>
<td></td>
</tr>
<tr>
<td>medPRR</td>
<td>4.75</td>
<td><strong>1.36</strong></td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>

| State B Data |
| meanSGP | 3.73 | **0.44** | 3.76 |
| medSGP | 4.04 | 3.79 | **1.33** |
| meanPRR | **0.61** | 4.17 | 3.76 |
| medPRR | 4.21 | **1.83** | 4.20 |

*Note.* The RMSDs for the Reading datasets are above the diagonal and those for the Mathematics datasets are below the diagonal. Bolded values indicate the RMSDs between the aggregated SGPs and PRRs that share a common aggregation function. meanSGP = mean Student Growth Percentile; medSGP = median Student Growth Percentile; meanPRR = mean Percentile Rank of Residual; medPRR = median Percentile Rank of Residual.
### Table 3

**Spearman Rank Order Correlations between each Pair of Aggregate-Level Conditional Status Metrics**

<table>
<thead>
<tr>
<th>Metric</th>
<th>meanSGP</th>
<th>medSGP</th>
<th>meanPRR</th>
<th>medPRR</th>
<th>FEM</th>
<th>REM</th>
<th>RMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>State A Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meanSGP</td>
<td>0.941</td>
<td><strong>0.997</strong></td>
<td>0.932</td>
<td></td>
<td>0.986</td>
<td>0.977</td>
<td>0.960</td>
</tr>
<tr>
<td>medSGP</td>
<td>0.965</td>
<td>0.938</td>
<td><strong>0.975</strong></td>
<td></td>
<td>0.903</td>
<td>0.900</td>
<td>0.887</td>
</tr>
<tr>
<td>meanPRR</td>
<td><strong>0.999</strong></td>
<td>0.964</td>
<td>0.935</td>
<td></td>
<td>0.988</td>
<td>0.980</td>
<td>0.963</td>
</tr>
<tr>
<td>medPRR</td>
<td>0.965</td>
<td><strong>0.994</strong></td>
<td>0.965</td>
<td></td>
<td>0.896</td>
<td>0.892</td>
<td>0.884</td>
</tr>
<tr>
<td>FEM</td>
<td>0.991</td>
<td>0.948</td>
<td>0.992</td>
<td>0.948</td>
<td></td>
<td>0.993</td>
<td>0.973</td>
</tr>
<tr>
<td>REM</td>
<td>0.987</td>
<td>0.943</td>
<td>0.988</td>
<td>0.944</td>
<td>0.996</td>
<td></td>
<td>0.970</td>
</tr>
<tr>
<td>RMR</td>
<td>0.928</td>
<td>0.880</td>
<td>0.933</td>
<td>0.885</td>
<td>0.937</td>
<td>0.937</td>
<td></td>
</tr>
</tbody>
</table>

|          |         |        |         |        |         |         |         |
| **State B Data** |         |        |         |        |         |         |         |
| meanSGP  | 0.959   | **0.997** | 0.953   |        | 0.989   | 0.985   | 0.905   |
| medSGP   | 0.986   | 0.956  | **0.989** |        | 0.933   | 0.930   | 0.857   |
| meanPRR  | **0.998** | 0.984  | 0.954   |        | 0.989   | 0.985   | 0.916   |
| medPRR   | 0.981   | **0.990** | 0.984   |        | 0.926   | 0.923   | 0.865   |
| FEM      | 0.993   | 0.976  | 0.995   | 0.975  |         | 0.995   | 0.900   |
| REM      | 0.992   | 0.976  | 0.995   | 0.975  | 0.999   |         | 0.899   |
| RMR      | 0.956   | 0.939  | 0.957   | 0.939  | 0.945   | 0.945   |         |

**Note.** The correlations for the Reading datasets are above the diagonals and those for the Mathematics datasets are below the diagonals. Bolded values in the upper left quadrant indicate the correlations between the aggregated SGPs and PRRs that share a common aggregation function. meanSGP = mean Student Growth Percentile; medSGP = median Student Growth Percentile; meanPRR = mean Percentile Rank of Residual; medPRR = median Percentile Rank of Residual; FEM = Fixed-Effect Metric; REM = Random-Effect Metric; RMR = Residuals from Mean Regression.
Figure Captions

Figure 1. Illustration of the contrast among the models used in deriving the various aggregate-level conditional status metrics using a $J = 1$ prior-grade empirical test score dataset. The student scores are expressed as conditional boxplots of current score on initial status and the school mean scores are overlaid as solid grey squares. PRRs are Percentile Ranks of Residuals; SGPs are Student Growth Percentiles; FEMs are Fixed-Effect Metrics; RMRs are Residuals from Mean Regression.

Figure 2. Illustration of the absolute differences in the percentile ranks of schools in the State B Reading data by their median Student Growth Percentiles (medSGPs; white boxplots) and Fixed-Effect Metrics (FEMs; grey boxplots) with each of the other aggregate-level conditional status metrics (medPRR = median Percentile Rank of Residuals; meanSGP = mean Student Growth Percentile; meanPRR = mean Percentile Rank of Residuals; REM = Random-Effects Metric; RMR = Residuals from Mean Regression).

Figure 3. Root mean square differences between aggregated Student Growth Percentiles (SGPs) and Percentile Ranks of Residuals (PRRs) from each pair of $J = 1, 2, 3$ prior-grade empirical datasets.

Figure 4. Average ranges of aggregate-level conditional status metrics in individual-level standard deviation units across transformations over schools or districts in each empirical dataset. From left, meanSGP = mean Student Growth Percentile, REM = Random-Effects Metric, RMR = Residuals from Mean Regression, FEM = Fixed-Effects Metric, meanPRR = mean Percentile Rank of Residual, medSGP = median Student Growth Percentile, and medPRR = median Percentile Rank of Residuals.