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

Katherine E. Castellano

University of California, Berkeley

Andrew D. Ho

Harvard Graduate School of Education

Author Note

KATHERINE CASTELLANO was an Institute of Education Sciences postdoctoral fellow at the University of California, Berkeley during the writing of this manuscript and is currently an Associate Psychometrician at the Educational Testing Service, San Francisco, CA; email: kcastellano001@ets.org. 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 Associate 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.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U. S. Department of Education, through grant R305B110017 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. The authors also benefited from the advice of Sophia Rabe-Hesketh of University of California, Berkeley. Any remaining errors are those of the authors alone.

Abstract

Aggregate-Level Conditional Status Metrics (ACSMs) describe the status of a group by referencing current performance to expectations given past scores. This paper provides a framework for these metrics, classifying them by aggregation function (mean or median), regression approach (linear mean, nonlinear quantile), and the scale supporting interpretations (percentile rank, score scale), among other factors. The study address the question, "how different are these ACSMs?" in three ways. First, using simulated data, it evaluates how well each model recovers its respective parameters. Second, using both simulated and empirical data, it illustrates practical differences among ACSMs in terms of pairwise rank differences incurred by switching between metrics. Third, it ranks metrics in terms of their robustness under scale transformations. The results consistently show that choices between mean- and median-based metrics lead to more substantial differences than choices between fixed- and random-effects or linear mean and nonlinear quantile regression. The findings set expectations for cross-metric comparability in realistic data scenarios.

Keywords: Student Growth Percentiles, Value-Added Models, group-level growth, conditional status

Practical Differences 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 or "difference from expectation" approach (Kolen, 2011). Following Castellano and Ho (2013a), at the individual level, we call these metrics "conditional status metrics" (CSMs), because they frame individual status in terms of conditional distributions given past scores. This class of metrics includes residuals from simple linear regression models (i.e., residual gain scores, Manning & DuBois, 1962). Another CSM is the 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, many important questions of policy and practice exist at the aggregate level. Current educational policies require longitudinal data summaries at the level of

classrooms, teachers, subgroups, schools, and states (U.S. Department of Education, 2005, 2009; Lissitz, Doran, Schafer, & Willhoft, 2006).

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 factors that differentiate among ACSMs, including the choice of aggregation function, nonlinear quantile vs. linear mean regression, and 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 aggregate-level SGP metrics and contrast them against alternatives, including aggregate-level Percentile Ranks of Residuals (PRRs; Castellano & Ho, 2013a) and simple approaches to so-called "value-added" models that attempt to isolate the causal effect of a teacher/school/principal on student achievement.

Within this latter 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). Instead, we aim to distinguish among the growing number of metrics and describe the practical differences of choices among them. We explicate differences among metrics in terms of 1) each model's recovery of its respective parameters, 2) practical differences among group percentile ranks for simulated and empirical data, and 3) robustness to scale transformations. Together, these analyses represent an effort to better understand the theoretical and practical differences among commonly used ACSMs, as well as the practical variability of any single ACSM under plausible alternative specifications.

Aggregate-Level Conditional Status Metrics

We begin by distinguishing models that support individual- and aggregate-level conditional status interpretations from two other classes of models that use longitudinal test

scores. Individual growth models (Rogosa & Willett, 1985; Singer & Willett, 2003) require test score data that share a common scale over time. Multivariate models are multivariate in their response variables and include 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). In contrast, CSMs do not require test score data that share a common scale over time, and they are univariate in their response variables.

McCaffrey et al. (2004) usefully describe individual growth models and CSMs as special cases of multivariate models, and we recommend their paper as a broader statistical framework. Castellano and Ho (2013b) contrast these models as they are used in state policies. A full discussion of "growth" and "value added" inferences is beyond the scope of this paper. We take the position that CSMs can inform discussions of 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

As previously stated, this paper presents three perspectives on differences among ACSMs: recovery of respective parameters, practical differences among ACSMs for simulated and real data, and robustness of each ACSM to scale transformations. We motivate each of these comparisons in this section.

Table 1 lists the nine ACSMs that we address in this paper and identifies distinguishing features along which these ACSMs differ. The nine ACSMs are median Student Growth Percentiles (medSGPs), mean SGPs (meanSGPs), median Percentile Ranks of Residuals (medPRRs), mean PRRs (meanPRRs), mean Residuals (meanResids), median Residuals (medResids), Residuals from Aggregate Regression (RARs), the Fixed-Effects Metric (FEM),

and the Random-Effects Metric (REM). Of the three distinguishing features, the first is the regression approach that establishes expectations for the target vector of test scores that, given they are typically from a recent timepoint, we refer to as the "current" scores. The models supporting ACSMs use different regression approaches: linear mean regression, including models that incorporate or disregard group membership (multilevel and single-level models, respectively), and nonlinear quantile regression. The second distinguishing feature is the aggregation function, where metrics support aggregate-level inferences through mean- or median-based operations. The third distinguishing 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, as residuals.

The theoretical and practical impact of these distinguishing principles has yet to be well described. Although there are bodies of literature supporting individual ACSMs, theoretical and empirical comparisons are often focused on examining the validity of value-added inferences of the various ACSMs as opposed to describing the magnitude of practical differences and the theoretical basis for cross-metric discrepancies. In particular, studies contrasting medSGPs against alternative ACSMs (e.g., Ehlert, Koedel, Parsons, & Podgursky, 2012; Goldhaber, Walch, & Gabele, 2012; Houng, & Justman, 2013; Guarino, Reckase, Stacy, & Wooldridge, 2014; Wright, 2010) are often critical of the use of median SGPs for value-added policies in comparison to other metrics. However, they have not identified the target parameter for median SGPs as we do here, and they place less emphasis on the basis for and magnitude of practical differences. Castellano and Ho (2013a) compare SGPs and PRRs at the individual level and find they are highly similar across a range of simulated and real data scenarios. We extend their work to the aggregate level and broaden the scope of metrics under consideration.

Finally, we compare ACSMs based on their robustness to scale transformations. This robustness is one of the motivations for the SGP metric (Betebenner, 2009). If SGPs are calculated for a matrix of student test score data, **X**, it is desirable to recover similar SGPs from a transformed matrix $T(\mathbf{X})$, where T represents a permissible monotonic transformation. Quantile regression is theoretically scale-invariant to monotonic transformations of at least the dependent variable (Koenker, 2005). Castellano and Ho (2012) found that for both simulated multivariate normal test score data and empirical data, individual SGPs were markedly more invariant to monotonic scale transformations than individual PRRs. With conditioning on one prior-test-score, they find that SGPs vary 1 to 2 percentile ranks of each other across the transformations, whereas student PRRs vary 4 to 5 percentile ranks on average. Our scale invariance study extends their results to consider aggregate-level metrics from Table 1.

Briggs and Betebenner (2009) compared the scale invariance of medSGPs and aggregate-level effects from a "layered," multivariate value-added model (Ballou, Sanders, & Wright, 2004). They used transformations that reflected linear or nonlinear growth, constant or increasing variance, and the more extreme exponential transformation. They found that conditioning on1, 2, and 3 prior-test-scores, medSGPs were near perfectly correlated across the transformations, whereas layered-model effects were less strongly correlated ($r > .9$), with smaller correlations arising from the exponential transformation ($r \approx .3$ to $.6$). Our scale invariance study differs from this study most importantly in our choice of scale transformations and the broader range of ACSMs that we consider.

The proliferating uses of ACSMs motivate a broad review of contrasting metrics as well as criteria for evaluating these metrics. We review each ACSM in Table 1, in turn, and then compare them in terms of parameter recovery, practical differences, and scale invariance.

Median and Mean Student Growth Percentiles

The SGP metric uses quantile regression to describe the "current" status of students in the context of their "prior" test performance. In practice, the "current" status is either the most recent set of test scores available or a time point that is of particular interest, and "prior" refers to score variables from one or more time points that precede the "current" time point. Castellano and Ho (2013a) describe the estimation of individual SGPs in detail and Betebenner (2010b) documents it in his "SGP" R (R Development Core Team, 2009) package. 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 contrasting it with linear mean regression of current scores on prior scores. In linear mean regression, this prediction takes the form of a conditional mean: an average current score for students given particular prior scores. In contrast, quantile regression allows for the estimation of conditional quantiles, such as the median, 25th percentile, and 90th percentile, for students with particular prior scores. Neither regression approach requires current and prior 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 observed score is between the .495 and .505 predicted quantile surfaces has an SGP of 50.

Just as linear mean regression can generalize to nonlinear functions, the SGP quantile approach employs nonlinear B-spline functions in fitting conditional quantiles to accommodate nonlinearity and heteroscedasticity in the data (Betebenner, 2009). We follow this approach,

implemented in Betebenner's (2010b) "SGP" package and thus classify SGPs as using "nonlinear quantile" regression. We also follow Castellano and Ho (2013a) 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 prevents cross-metric comparisons from being confounded by decisions about rounding.

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. 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 equal-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. We consider threats to interpretations of ordinal-based averages as a matter of degree. We demonstrate that these threats may be offset by the advantages of means.

Percentile Ranks of Residuals

Castellano and Ho (2013a) contrast SGPs with an analogous approach that uses the percentile ranks of residuals (PRRs) from the linear mean 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 (2013a) 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 "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 they are easily aggregated to meanPRRs and medPRRs. PRRs are computed by first regressing students' current scores (*Y*) on their scores from *J* prior time points ($X_1, X_2, \ldots, X_J$) as follows, where *i* indexes student and *g* indexes group:

$$Y_{ig} = \alpha + \sum_{j=1}^{J} \beta_j X_{jig} + \epsilon_{ig}. \quad (1)$$

Here, α is the intercept, the β_j parameters are the regression coefficients for each prior test score included, and $\epsilon_{ig}$is the individual error term. Equation (1) does not take into account group membership; we will contrast it with upcoming models that do. As a linear model for the conditional mean, this approach also contrasts with the quantile regression model underlying SGPs. The operationally-used SGP quantile regressions are nonlinear, and there are typically 100 quantile regression functions (and in our case, 1000) instead of one linear regression function. The mean regression framework easily accommodates nonlinear specifications of the prior scores (Akram, Erickson, & Meyer, 2013). For parsimony, we only include linear mean specifications in this paper.

The residuals for the regression shown in Equation (1) are the simple differences between the observed and expected values given past scores: $e_{ig} = y_{ig} - \hat{y}_{ig}$. The PRRs are the residuals

e_{ig} for student *i* in group *g*, rank ordered and transformed to percentile ranks, which we express as $\text{PRR}_{ig} = \text{F}(e_{\text{ig}}) \times 100$, where F( ) is the empirical cumulative distribution function. We round the PRRs to one decimal point for consistency with the rounding of SGPs. Although studentized residuals could be used, we follow Castellano and Ho (2013a) and use raw residuals for simplicity and because differences are trivial, particularly at the aggregate level. We aggregate PRRs over students within groups using the mean and median function to obtain meanPRRs and medPRRs respectively, for example, $\text{meanPRR}_g = \frac{\sum_{i=1}^{n_g} F(e_{ig})}{n_g}$, where n_g is the number of students in group *g*.

Mean and Median Residuals

Another operationally simple approach involves the aggregation of raw residuals derived from Equation (1) without performing the percentile rank transformation. In contrast to meanPRR_g, $\text{meanResid}_g = \frac{\sum_{i=1}^{n_g} e_{ig}}{n_g}$. Mean residuals have some precedent in the value-added literature as they, or "precision-weighted" mean residuals, have been used as value-added metrics (e.g., Kane & Staiger, 2008; Chettey, Freidman, & Rockoff, 2011; Guarino, Reckase, & Wooldridge, 2012). We do not include the medResid metric any further than its listing in Table 1, because it selects the same mid-ranked student in any group as do medPRRs. Although they differ in a nonlinear fashion along their scales for interpretation, the two metrics will be perfectly rank-correlated if all groups have an odd number of students and nearly perfectly rank-correlated otherwise. MeanResids have no such easily specified relationship with other metrics, thus we retain them for comparison.

Residuals from Aggregate Regression

The SGP and PRR metrics first condition and then aggregate in two distinct steps. The

RAR metric, in contrast, results from fitting a regression model to aggregate values, reversing the order of operations to aggregating first and then conditioning. The RAR metric thus represents the conditional status of aggregates instead of an aggregate of conditional status. As the name suggests, RARs are the residuals from the regression of average current scores on their *J* prior 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). The RAR metric can use the same linear mean regression as the PRR metric, but the individual scores in Equation 1 are replaced by their group averages, reducing the fitted data from *n* students to *G* groups:

$$\bar{Y}_{.g} = \alpha^{(B)} + \sum_{j=1}^{J} \beta_j^{(B)} \bar{X}_{j.g} + \epsilon_g \quad (2)$$

Here, $\bar{Y}_{.g}$ is the mean current score for group g, $\bar{X}_{j.g}$ represents the g^{th} group's mean scores for the t^{th} prior time-point, $\beta_j^{(B)}$ refers to the corresponding regression coefficients from the "between-groups" regression, $\alpha^{(B)}$ denotes the intercept, and ϵ_g is group g's error term. The RAR is the difference between a group's observed and expected current average score ($\bar{y}_{.g} - \hat{\bar{y}}_{.g}$).

Fixed-Effects Metric

The model supporting FEMs is sometimes described as a "covariate adjustment" model or, more simply, as an Analysis of Covariance (ANCOVA). In this approach, an aggregate-level "fixed effect" is added to the individual-level linear mean regression of current scores on *J* prior scores. The fixed effects can be represented by the coefficients (γ_g) of dummy variables for groups, resulting in distinct intercepts for each group g:

$$Y_{ig} = \gamma_g + \sum_{j=1}^{J} \beta_j^{(W)} X_{jig} + \epsilon_{ig}. \quad (3)$$

Each group's FEM is operationalized as its estimate of γ_g with respect to a reference group or subject to another constraint such as $\sum_g \gamma_g = 0$, which we use. 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 a linear mean regression framework.

The multilevel modeling literature provides many useful links between models. As one illustrative contrast, when group sizes are equal and there is one prior grade predictor (X), the overall regression coefficient, β, is the weighted sum of the between- ($\beta^{(B)}$) and within-group ($\beta^{(W)}$) regression coefficients, $\beta = \eta_X^2\beta^{(B)} + (1 - \eta_X^2)\beta^{(W)}$, where η_X^2 is the "correlation ratio," an intraclass correlation coefficient for X that is precision-adjusted by dividing by the reliability of its group mean (Snijders & Bosker, 2011, p. 31). When this unconditional intraclass correlation is low, that is, when there is substantially more within-group than between-group variation on X, the regression coefficients that determine meanPRRs and meanResids (β) will be similar to the coefficients that determine FEMs $\left(\beta^{(W)}\right)$.

The focus of this paper is not on regression coefficients but the contrasts between ACSMs. However, there is reason to expect that similarities in slopes will lead to similarities between ACSMs, particularly between FEMs and meanResids. The FEMs can in fact be expressed as a mean residual from a reference line with slope $\beta^{(W)}$: $\mathrm{FEM_g} = \bar{y}_{.g} - \sum_{j=1}^{J} \beta_j^{(W)} \bar{x}_{j.g}$. Thus, when unconditional intraclass correlations are low, we can expect not only that β and $\beta^{(W)}$ will be similar but also that FEMs and meanResids will be highly correlated.

Random-Effects Metric

The model supporting the REM is a two-level "random intercept" model or a random-effects ANCOVA (Raudenbush & Bryk, 2002). It is similar to the model supporting FEMs, but

random-intercept models do not parameterize the intercepts directly; rather, they treat them as random variables. Given school-level random intercepts designated as u_g and an overall average group intercept designated as γ_0:

$$Y_{\text{ig}} = (\gamma_0 + u_g) + \sum_{j=1}^{J} \beta_j^{(R)} X_{jig} + \epsilon_{\text{ig}}, \quad (4)$$

This random-intercepts model also constrains all groups to have the same slope estimates ($\hat{\beta}_j^{(R)}$). Rabe-Hesketh and Skrondal (2012) explain that this model implicitly assumes that $\beta_j^{(W)}$ and $\beta_j^{(B)}$ are equal. Under a feasible generalized least squares estimation procedure, they express the random effects estimate of the slope as a precision-weighted average of $\hat{\beta}_j^{(W)}$ and $\hat{\beta}_j^{(B)}$ (p. 148). Their expression implies that $\hat{\beta}_j^{(B)}$ will be down-weighted, and the coefficients underlying the REM ($\hat{\beta}_j^{(R)}$) and FEM ($\hat{\beta}_j^{(W)}$) metrics will be similar when 1) the variance of the random intercepts is large with respect to the level-1 error variance and 2) within-group variances of predictors are large with respect to their between-group variances. These imply large conditional intraclass correlations for the outcome and small unconditional intraclass correlations for the predictors, respectively. Notably, this latter criterion, low unconditional intraclass correlations for predictors, is common for educational test scores nested within schools (Hedges & Hedberg, 2007).

Even when slopes are similar, FEMs and REMs may differ due to the necessary estimation of each REM (u_g). We follow common practice (Rabe-Hesketh & Skrondal, 2012) and shrink REMs to the overall mean using Empirical Bayes estimation. 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. In the practical

scenarios that we will illustrate, we found that the shrunken and unshrunken estimates were almost perfectly correlated and thus do not consider unshrunken estimates further.

Visual Comparison of ACSMs

Figure 1 contrasts the models that support ACSMs by illustrating their fit to a scatterplot of real data. Light gray conditional boxplots show the empirical bivariate distribution of a statewide cohort's Grade 6 scores given each score point in their prior grade. The linear mean regression line in dark gray has a slope $\hat{\beta}_1$ (Equation 1). It serves as a reference for meanResids and PRRs. Points above the line represent students with positive residuals and high PRRs who performed better in Grade 6 than expected given their Grade 5 scores. Groups with more students above the line will receive higher meanResids and aggregated-PRRs than groups with more students below the line.

To illustrate SGP operationalization, Figure 1 also shows the median quantile spline in black. (It is actually the line for the .495 quantile, and scores falling between the .495 and .505 quantile lines receive an SGP of 50.) Like PRRs, points above the line will be assigned higher SGPs. The jagged shape of the spline at extreme score points arises from corrections that ensure that quantile regression lines do not cross each other (Betebenner, 2010b; Castellano & Ho, 2012; Dette & Volgushev, 2008).

The dashed line denotes the regression line for the RARs 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 RAR line is steeper compared to the reference lines of the other models. Because RARs ignore within-group variability, and because within-group variability far exceeds between-group variability in educational test score data (Hedges & Hedberg, 2007), we expect

and ultimately find that RARs will produce the most distinct rank orderings of groups than any of the other ACSMs. For parsimony, and because ignoring within-group variation with RARs is poorly motivated in practice, we exclude RARs from the remainder of the paper.

The fixed effects model supporting FEMs effectively estimates a common slope for within-group data, $\hat{\beta}_1^{(W)}$ (Equation 3). Figure 1 shows this dotted regression line when $\gamma_g = 0$, under the constraint that intercepts sum to 0. All groups have parallel regression lines (not shown), and vertical distances between the regression line of any group and that of the dotted line is that group's FEM. The line for the REM model differs but is visually indistinguishable from the FEM line for these data. The REM model produces an overall regression function with slope $\hat{\beta}_1^{(R)}$ (Equation 4), following which estimation of REMs ($\hat{u}_g$) for each group proceeds by Empirical Bayes estimation.

The total (resids, PRRs), within-group (FEMs), and between-group (RARs) regression lines are visually distinguishable. As the previous sections described, the total regression coefficient will be closer to the within-group coefficient when between-group variation on X is relatively low, that is, low intraclass correlations for predictors. This holds in Figure 1, where the slope of the linear mean regression line is closer to the within-group line than the between-group line. The relative dominance of within-group variation in the predictors also explains our inability to distinguish between the fixed- and random-effects lines visually. These similarities motivate the theoretical and empirical analyses that follow.

ACSM Parameter Recovery

We first evaluate each ACSM according to how well it recovers its corresponding parameter or expected value. As described in the preceding sections, each ACSM is derived under a different model and/or computations with different parameters or intended targets. In

this section, we describe our data generating procedure (DGP), define the parameters for each ACSM under this DGP, and then evaluate the recovery of each ACSM using bias, Root Mean Square Error (RMSE), and correlations with their respective true values.

Data Generating Process

The ideal DGP allows control over parameters that ensure realistic relationships among variables while also defining parameters for evaluating ACSMs. We considered generating data using the random-intercept model as given in Equation 4. However, such a DGP would privilege the random-effects metric over the other ACSMs; that is, the other ACSMs do not have parameters readily derivable from parameters specified in Equation 4. Moreover, such an approach precludes specification of the between-group correlations between each respective predictor and the outcome. In the context of year-to-year test scores, in which both the predictors and outcome are test scores, it is not realistic to have an outcome variable that has correlations with predictors substantially different than the correlations among the predictors themselves.

We therefore use a "decomposed multivariate normal" DGP that generates data as the sum of between-group and within-group multivariate normal (MVN) distributions. The between-group data are the group means, and the within-group data are individual deviations from the group means, or the group-mean-centered values. We use four total years of test scores—one "current grade" and three "prior grades" —following the common practice of using ACSMs when there are at least three prior timepoints (e.g., Sanders, 2006). The generating model is thus:

$$\begin{pmatrix} Y_g^B \\ X_{1g}^B \\ X_{2g}^B \\ X_{3g}^B \end{pmatrix} \equiv \begin{pmatrix} Y_g^B \\ \boldsymbol{X_g^B} \end{pmatrix} \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma^B}) \qquad \begin{pmatrix} Y_{ig}^W \\ X_{1ig}^W \\ X_{2ig}^W \\ X_{3ig}^W \end{pmatrix} \equiv \begin{pmatrix} Y_{ig}^W \\ \boldsymbol{X_{ig}^W} \end{pmatrix} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma^W})$$

$$Y_{ig} = Y_g^B + Y_{ig}^W \qquad \boldsymbol{X} = \boldsymbol{X_g^B} + \boldsymbol{X_{ig}^W}.$$

Here, Y_{ig} denotes the nominal current test score for student *i* in group *g*, and $\boldsymbol{X}$ is a $N \times J$ matrix of student test scores, where $N = \sum_{g=1}^{G} n_g$ and $J = 3$ denotes the number of prior years.

We define control parameters to align with the simulation specifications of Castellano and Ho (2013) that reflect patterns observed in real data. The mean vector is $\boldsymbol{\mu} = [\mu_Y \ \boldsymbol{\mu_X}]' = [265, 260, 255, 250]'$. The covariance matrices $\boldsymbol{\Sigma^B}$ and $\boldsymbol{\Sigma^W}$ are expressed in terms of an overall, or as previously termed "total," correlation matrix, $\boldsymbol{R}$ (which we denote without a superscript), a between-group correlation matrix, $\boldsymbol{R^B}$, a diagonalized matrix of standard deviations, $\boldsymbol{D}$, and an intraclass correlation coefficient, ω, as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{DRD}$$

$$\boldsymbol{\Sigma^B} = \left(\sqrt{\omega}\boldsymbol{D}\right)\boldsymbol{R^B}\left(\sqrt{\omega}\boldsymbol{D}\right)$$

$$\boldsymbol{\Sigma^W} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma^B}.$$

Following Castellano and Ho (2013), for $\boldsymbol{R}$, we set correlations between adjacent year variables (lag-1 correlations) to .78, lag-2 correlations to .76, and lag-3 correlations to .74. Informed by patterns in our real data, we set corresponding lagged pairwise correlations between group means in $\boldsymbol{R^B}$ to .90, .85, and .80. We set standard deviations in $\boldsymbol{D}$ constant at $\sqrt{65}$. The intraclass correlation, ω, controls the amount of non-random sorting into groups, and we vary this as a factor, from .05 to .15 to .25. These values span a range observed in practice (Hedges and Hedberg, 2007) and in our own empirical datasets (from .06 to .15).

To mimic a reasonable number of schools observed in practice, we assign students to 500 groups with equal group sizes. We investigate three group sizes, n_g =25, 50, and 100. Each of the nine crossed combinations of group sizes (25, 50, and 100) by unconditional ICC (.05, .15, and .25) was replicated 100 times.

Expected Values for Each ACSM

In this section, for each simulation condition, we derive the expected values for each ACSM, beginning by establishing properties of group average residuals (meanResids), proceeding to aggregates of transformed residuals (aggregated SGPs/PRRs), and concluding with FEMs. Under random allocation of students to teachers, SGPs and PRRs are uniformly distributed 0 to 100 for each group, resulting in expected values of 50 for mean and median SGPs and PRRs. Under nonrandom allocation of students to teachers, such as in our DGP controlled by nonzero ICCs, we can also derive the expected values of group residuals and thereby mean and median SGPs and PRRs.

We first consider the conditional distribution of the current scores given the prior scores. Under our decomposed MVN DGP, this distribution has known variance σ_ϵ^2:

$$\sigma_\epsilon^2 = \sigma_{Y|X}^2 = \sigma_Y^2 - \boldsymbol{\Sigma_{YX}}(\boldsymbol{\Sigma_{XX}})^{-1}\boldsymbol{\Sigma_{Y'X}},$$

where $\boldsymbol{\Sigma_{YX}}$ is a column vector of covariances between Y and each prior score X_1, X_2, and X_3, and $\boldsymbol{\Sigma_{XX}}$ is the covariance matrix among the prior grade-level scores in $\boldsymbol{X}$. The normality of the population conditional distribution allows us to derive the meanPRR and medPRR as a closed-form expression. First, note that under our decomposed MVN DGP, the unconditional distribution of residuals over all groups is:

$$\epsilon \sim N(0, \sigma_\epsilon^2).$$

We denote the residuals for students within, or given, a particular group g using conditional notation, $\epsilon|g$. By construction, these are normally distributed:

$$\begin{aligned} \epsilon|g &\sim N\left(\mu_{\epsilon|g}, \sigma^2_{\epsilon|g}\right) \\ \mu_{\epsilon|g} &= Y^B_g - \mu_{Y|X} = Y^B_g - \left[\mu_Y + \boldsymbol{\Sigma_{YX}}(\boldsymbol{\Sigma_{XX}})^{-1}\left(\boldsymbol{X^B_g} - \boldsymbol{\mu_X}\right)\right] \\ \sigma^2_{\epsilon|g} &= \sigma^{2(W)}_{Y|X} = \sigma^{2(W)}_{Y} - \boldsymbol{\Sigma^W_{YX}}\left(\boldsymbol{\Sigma^W_{XX}}\right)^{-1}\boldsymbol{\Sigma^W_{Y'X}}. \end{aligned} \tag{5}$$

The mean of the residuals within group g, $\mu_{\epsilon|g}$, represents the average expected residual over students in group g from the student-level (overall) regression in Equation 1. These are thus the true group meanResids, $\text{meanResid}_g^{\text{true}}$.

Due to the normality of overall residuals ϵ, the PRRs are the normal CDF of these standardized residuals: $PRR = \Phi\left(\frac{\epsilon}{\sigma_\epsilon}\right)$, where Φ denotes the standard normal CDF. The percentile rank of a residual from any group is therefore: $PRR|g = \Phi\left(\frac{\epsilon|g}{\sigma_\epsilon}\right)$, with $\epsilon|g$ distributed as specified in Equation 5. Accordingly, the expected value of $PRR|g$ for a particular group can be expressed in closed form as a normal CDF of a normal deviate:

$$\text{meanPRR}_g^{true} = \Phi\left(\frac{\mu_{\epsilon|g}}{\sqrt{\sigma^2_\epsilon + \sigma^2_{\epsilon|g}}}\right).$$

Additionally, because the distribution of residuals is normal, the parameters of the conditional quantiles follow those of a normal distribution. The associated conditional percentile ranks are the definition of SGPs and are equal to $\Phi\left(\frac{\epsilon}{\sigma_\epsilon}\right)$. Thus, under our DGP, population SGPs are equal to population PRRs, and $\text{meanSGP}_g^{true} = \text{meanPRR}_g^{true}$.

The expected value of medPRRs and medSGPs follow from the symmetry of the normally distributed residuals for each group, such that the median residual equals the

mean: $\mu_{\epsilon|g}$. The Φ transformation is monotonic, ensuring that the median transformed value is the transformed median value. The medPRR parameter is therefore:

$$\text{medPRR}_g^{true} = \Phi\left(\frac{\mu_{\epsilon|g}}{\sigma_\epsilon}\right).$$

As before, the normality of the conditional distribution results in $\text{medSGP}_g^{true} = \text{medPRR}_g^{true}$.

The expected values for FEMs, like those for meanResids in Equation 5, can be expressed in terms of population regression coefficients implied by the DGP. That is, $\boldsymbol{\beta} = \boldsymbol{\Sigma_{XX}^{-1}\Sigma_{YX}}$, and $\boldsymbol{\beta^W} = \left(\boldsymbol{\Sigma_{XX}^W}\right)^{\boldsymbol{-1}}\boldsymbol{\Sigma_{YX}^W}$. We then compute the groups' expected residuals by subtracting the generated group means from the expected group means:

$$\text{FEM}_\text{g}^\text{true} = Y_g^B - \mu_{Y|X}^W = Y_g^B - \left[\mu_Y + \boldsymbol{\Sigma_{YX}^W}\left(\boldsymbol{\Sigma_{XX}^W}\right)^{\boldsymbol{-1}}\left(\boldsymbol{X_g^B} - \boldsymbol{\mu_X}\right)\right]$$

All of these terms are specified or generated as part of the DGP. The population means, μ_Y and $\boldsymbol{\mu_X}$, represent subsets of the specified mean vector $\boldsymbol{\mu}$ for the current and prior scores, respectively.

This DGP does not allow for straightforward expressions of $\boldsymbol{\beta^R}$ so we do not include parameter recovery for REMs. However, given the relatively low magnitudes of the ICCs for the prior grade levels, there is relatively little between-group variability in the conditioning variables, thus $\boldsymbol{\beta^R} \approx \boldsymbol{\beta^W}$ **(**Rabe-Hesketh and Skrondal, 2012, p. 148). Although we do not present REM recovery of FEM parameters, we note that the results are nearly identical to the presented FEM recovery of FEM parameters. Equivalently, as we do show, REM and FEM metrics are nearly perfectly correlated in our simulated and empirical data examples.

ACSM Recovery of Expected Values

We evaluate each estimated ACSM ($\hat{\theta}$) on its recovery of its respective expected value (θ) using bias, RMSE, and correlations, which we compute by averaging over groups and replications for a given condition:

$$Bias(\hat{\theta}) = \frac{1}{R \times G}\sum_{r=1}^{R}\sum_{g=1}^{G}(\hat{\theta}_{gr} - \theta_{gr}),$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{R \times G}\sum_{r=1}^{R}\sum_{g=1}^{G}\left(\hat{\theta}_{gr} - \theta_{gr}\right)^2},$$

$$Corr\left(\theta, \hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\frac{\sum_{g=1}^{G}\left[\left(\hat{\theta}_{gr} - \bar{\hat{\theta}}_r\right)\left(\theta_{gr} - \bar{\theta}_r\right)\right]}{\sqrt{\sum_{g=1}^{G}\left(\hat{\theta}_{gr} - \bar{\hat{\theta}}_r\right)^2 \sum_{g=1}^{G}\left(\theta_{gr} - \bar{\theta}_r\right)^2}}.$$

Here, *r* denotes replication with *R*=100 and *g* denotes group with *G*=500. Bias and RMSE are expressed on the scale of the ACSM of interest and are thus not comparable across all ACSMs. Specifically, these values should not be compared between the residual-based metrics, FEMs and meanResids, and the percentile-rank-based metrics, mean/medSGPs and mean/medPRRs. To allow for comparison across all metrics, we also report Pearson correlations between each estimated ACSM and its expected value, averaged over the 100 replications for each condition. Table 2 gives these RMSEs and Pearson correlations. All of the biases were essentially zero; thus, we do not report them in the interest of space. Note that RMSEs are not strictly standard errors for an estimate of a single parameter as each group has its own expected value. Rather, the RSMEs are average standard errors over the group parameters generated by the DGP.

First, we review the performance of the percentile-rank scaled metrics. From Table 2, it is apparent that the median SGPs and PRRs underperform their mean-based counterparts in terms of parameter recovery. The RMSEs for the median-based percentile rank metrics are consistently about 1.6 times larger than those of the corresponding mean-based metrics across all simulation conditions. Both meanSGPs/PRRs and medSGPs/PRRs are unbiased, but meanSGPs/PRRs are substantially more efficient. Moreover, results for SGPs and PRRs are

almost identical, indicating that under multivariate normality of cross-year test scores, the choice between conditional quantile and conditional mean regression is immaterial for aggregate-level conditional status interpretations.

For the two residual scaled metrics, FEMs and meanResids, we find that their average efficiency is indistinguishable under these simulation conditions. Recall that FEMs are average residuals from a within-group regression (with $\boldsymbol{\beta^W}$) and meanResids are average residuals from an "overall" student-level regression (with $\boldsymbol{\beta}$). Thus, the nearly identical RMSEs for FEMs and meanResids in Table 2 are partly anticipated by the expected similarity of $\boldsymbol{\beta}$ and $\boldsymbol{\beta^W}$ when intraclass correlations are low, although the similarity is striking even when intraclass correlations are 0.25. The consideration of group membership in the estimation of regression slopes and residuals for FEMs buys little precision under these simulation parameters. This foreshadows upcoming results as we shift from comparing efficiency of metrics to comparing their magnitudes and rank orderings for individual groups: there is considerable similarity across meanResids, FEMs, and REMs for both empirical and simulated datasets.

We express the RMSEs for FEMs and meanResids in Table 2 in terms of standard deviation units of Y ($\sigma_Y = \sqrt{65}$) rather than on the original Y scale to facilitate interpretability. The RMSEs are between 1/9 and 1/20 of a standard deviation unit on this scale, and RMSEs do not seem to depend greatly upon ICCs. An alternative scale for interpreting RMSEs is the standard deviation of $\text{meanResid}_{\text{g}}^{\text{true}}$, $\text{SD}(\mu_{\epsilon|g})$, which for our DGP equal $\left(\sigma_Y^{2(B)} + \boldsymbol{\beta\Sigma_{XX}^{B}\beta} - 2\boldsymbol{\beta\Sigma_{YX}^{B}}\right)^{.5}$ These RMSEs are displayed in the last column of Table 2 for both FEMs and meanResids, given that they are almost identical. On this scale, RMSEs are over a standard deviation unit when $n_g = 25$ and $\omega = .05$. It is only with sample sizes of 50 and ICCs of .15 or higher that RMSEs drop below 1/2 of a standard deviation. The differences among RMSEs

across ICCs are also more pronounced for FEMs and meanResids when RMSEs are expressed on the scale of $\text{meanResid}_g^{\text{true}}$. These differences are consistent with the differences across ICCs in the correlation results at the bottom of Table 2. Estimates are more efficient as ICCs rise, moderately on the σ_Y scale and substantially on the $\text{SD}(\mu_{\epsilon|g})$ scale.

Across all metrics, we observe that the two varied simulation factors—group size and grade-level ICC—affect recovery similarly. Each metric's recovery of its expected value is better for higher unconditional ICCs and larger group sizes as observed by the lower RMSEs and higher Pearson correlations under these conditions.

To clearly compare the performance across all metrics, Figure 2 illustrates the mean correlations for the smallest group size, $n_g = 25$, as this level shows the largest differences among the metrics. As Table 2 shows the same general pattern holds for group sizes of 50 and 100 as well. Figure 2 clearly shows that the median-based percentile rank metrics (medPRR, medSGP) demonstrate the poorest recovery, the mean-based percentile-rank metrics (meanPRR, meanSGP) demonstrate better recovery, and the mean-based residual metrics (meanResid, FEM) show the best recovery. The meanSGP/PRR correlations are more similar to those for the meanResids and FEMs than to those for the medSGPs/PRRs. In summary, the efficiency of the medSGPs/PRRs is noticeably poor compared to the performance of all the other ACSMs.

Theoretical Differences between Mean and Median Percentile Rank ACSMs

In this section, we use theoretical statistical results to explain the substantial differences between mean-based and median-based percentile rank metrics observed in Table 2 and Figure 2. As percentile ranks, SGPs and PRRs follow a theoretical uniform distribution with a 0 to 100 range and thus have a standard deviation of $50/\sqrt{3}$. When there are no average differences across groups ($\omega = 0$), the sampling distribution of medians for group sizes of n is that for the

$(n+1)/2$ order statistic and is known to be distributed as a beta distribution with parameters $\alpha = \beta = (n+1)/2$ (Casella & Berger, 2001).

The expected value of medSGPs and medPRRs under these conditions is 50, and their standard error is $50/\sqrt{n+2}$. The expected value for the meanSGPs and meanPRRs is also 50, but their standard error is $50/\sqrt{3n}$. 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 square of this factor, 3, is the Asymptotic Relative Efficiency (ARE) of means over medians. Under random assignment, a median PRR or SGP will require three times the sample size to achieve the same efficiency as a mean PRR or SGP.

When there are average group differences, the sampling variability of mean and median percentile ranks, like the expected value, depend upon the variance of the unconditional residuals, σ_ϵ^2, and the conditional mean and variance of group residuals $\mu_{\epsilon|g}$and $\sigma_{\epsilon|g}^2$, respectively. In Figure 3, we use Monte Carlo simulations to display the standard errors of median (dashed lines) versus mean (solid lines) PRRs and SGPs under the decomposed MVN DGP for $n_g = 25$ (black lines) and $n_g = 100$ (grey lines) when $\omega = .15$. By plotting the standard errors (Figure 3a) and relative efficiencies (Figure 3b) for metrics on the location of $\text{meanResid}_g^{\text{true}}$ in terms of their standard deviation, $\text{SD}(\mu_{\epsilon|g})$, Figure 3 shows that these statistics are lower for groups on either extreme of the distribution. Estimates are more precise and means and medians are more similar in efficiency when true group mean residuals are extreme. On average, however, the relative efficiency of means over medians is considerable across a range of realistic group mean values. Figure 3b shows that, on average (at ± 1 SD, given the underlying standard normal density), medians require at least 2.5 times the sample size of means to achieve

the same efficiency. The magnitude of these differences in sampling variability represents a strong argument for using mean-based SGPs or PRRs over their median counterparts.

Empirical Data

To provide a real-data perspective on these findings, we use the same two statewide data files as those from the real-data analyses of Castellano and Ho (2013), allowing for a common data reference point between these papers. There are a total of four distinct four-year longitudinal datasets: two states with data in each of two subjects. The states, referred to as "State A" and "State B," contrast usefully in size and scaling procedures. The State A dataset contains records for a single cohort of about 25,000 students with reading and mathematics scores from grade 3 to grade 6 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 to grade 6 on within-grade-scaled tests.

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 the purposes of these illustrative analyses, students with missing data are excluded. For convenience, we describe district- and school-level grade cohorts as "groups." To adhere to standard reporting rules, we follow Colorado's cutoff for reporting medSGPs (CDE, 2012) and exclude results for groups 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 groups from aggregate-level analyses. In contrast, FEMs and REMs use aggregate-level information early in estimation; thus, for these metrics, we exclude students from small groups prior to any calculations.

This decision rule excludes about 18 percent of districts (2.8 percent of students) in State A and about 20 percent of schools (1 percent of students) in State B. In State A, there are 272 districts for Reading and 273 for Math. The median and mean group sizes for each subject are 46 and 90 respectively, with a maximum size of 1700. In State B, there are 546 schools in the Reading dataset and 542 in the Math dataset. The median and mean group sizes are 133 and 139 respectively, with a maximum size of 380.

Cross-Metric Comparability

In the previous section, we compared ACSM's in the recovery of their own respective parameters, whereas in this section, we compare metrics to each other directly for both simulated and empirical data. We use rank-based methods to compare the consequences of switching among metrics because of the nonlinear relationship between the residual scale and the percentile rank scale. Ranks are also a useful scale to communicate practical differences. Table 3 displays Spearman rank correlations for simulated and empirical data, and Figure 4 shows empirical distributions of percentile rank differences for groups.

Correlations among ACSMs

The top third of Table 3 gives average correlations over replications for simulated data when n_g = 25 and ICCs=0.05 (above the diagonal) and n_g = 25 and ICCs=0.25 (below the diagonal). We chose the smallest group size condition, as it shows the starkest contrast among the metrics and reflects typical class sizes. Thus, these correlations represent possible correlations between ACSMs at the teacher level. The bottom two thirds of Table 3 show correlations for the empirical data in States A and B, in reading and mathematics. We divide the ACSMs into mean-based residual metrics, mean-based percentile rank metrics, and median-based percentile rank metrics with a 3x3 "tic-tac-toe" grid in each matrix.

We note three primary observations from the simulated and empirical results in Table 3. First, from the upper left of these 3x3 grids, we see that REMs, FEMs, and meanResids are nearly perfectly correlated with each other. This holds even for the simulated data with relatively small group sizes of 25. Second, meanSGPs and meanPRRs are also nearly perfectly correlated with each other and, to a slightly lesser degree, to the mean-based residual metrics. Third, medSGPs and medPRRs are highly correlated with each other but are much less correlated with the other metrics.

The high correlations between REMs, FEMs, and meanResids are predictable from Figure 1, where the three coefficients $\hat{\beta}_j$, $\hat{\beta}_j^{(R)}$, and $\hat{\beta}_j^{(W)}$, are almost impossible to distinguish visually. The low unconditional ICCs, both in the simulations and in the real data, allow within-group information to dominate and bring $\hat{\beta}_j$ and $\hat{\beta}_j^{(R)}$ closer to $\hat{\beta}_j^{(W)}$. The results show that this similarity between coefficients propagates to their aggregate-level residuals. The magnitudes may differ slightly, particularly when shrunken REM estimates are used. However, our results show that the rank ordering of the groups will not change considerably, due in part to the more or less balanced group sizes. This study shows that for education data, for which there generally is much more within-group variation than between-group variation, the choice between fixed- and random-effects may be immaterial, at least when only regressing on prior scores.

The strong similarities between meanSGPs and meanPRRs are aggregate-level extensions of the findings of Castellano and Ho (2013). Table 3 further shows that these two metrics are highly correlated with the mean-based residual metrics. In the empirical data, the correlations between mean-based residual and percentile rank metrics are highest for State B math. Comparing these correlations for the simulated data, above the diagonal (ICC of .05) and below the diagonal (ICC of .25), we can see that these correlations will be higher when unconditional

ICCs are higher. Indeed, the State B math data had the highest unconditional ICCs of the four empirical datasets. Additionally, State B has larger group sizes than State A. This is consistent with simulated results that show higher and more similar correlations among metrics when group sizes are larger.

The median-based metrics are the most different from the other metrics. Correlations between the medSGPs and medPRRs with the FEMs, REMs, and meanResids are much lower than those between the meanSGPs and meanPRRs. 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. Notably, Table 3 shows that correlations between SGP-based metrics and PRR-based metrics are much higher than correlations between mean-based and median-based metrics. The relative impact of the choice of aggregation function is far greater than the relative impact of the choice of regression model.

Absolute Differences in Group 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 an interpretable scale. These differences in percentile ranks should not be confused with differences in aggregated SGPs and PRRs that are themselves on the percentile rank scale. As an example, the medSGP of one group is 32, which is greater than or equal to the medSGPs of 36% of all the groups. This group is thus at the 36th percentile on the medSGP metric. The same group has a

meanSGP of 40, which corresponds to the 26th percentile on the meanSGP metric. For this group, switching metrics leads to an absolute difference in percentile ranks of $|36 - 26| = 10$.

Figure 4 uses boxplots to show the distributions of absolute percentile rank differences between ACSMs and two "reference metrics," (a) medSGPs because they are widely used in practice, and (b) FEMs because they have the best recovery of their expected values (see Figure 2). Simulated results are on the left, and empirical results are on the right. 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. The simulated results on the left use a single randomly selected replication with group sizes of 100 and grade-level ICCs of .05. These correspond most closely with the group sizes and ICCs of the State B reading data on the right. Any of the other empirical datasets could have been used to illustrate the same findings.

Like Table 3, Figure 4 shows that comparisons of the metrics for the simulated and empirical data are very similar, suggesting that our simulation modeled the real data well for ACSM purposes. The rank differences between medSGPs and mean-based metrics is large in both simulated and empirical data (Figure 4a). The rank differences between FEMs and other mean-based metrics are lower for simulated than for empirical data (Figure 4b). Consistent with the results in Table 3, in Figure 4, the ACSMs fall into three distinct categories for both the simulated and empirical data: mean-based residual metrics (REM, FEM, meanResid), mean-based percentile rank metrics (meanSGP, meanPRR), and median-based percentile rank metrics (medSGP, medPRR).

Figure 4 supplements Table 3 with a finer grain picture of the comparability of the ACSMs. The small number to the left of each boxplot refers to the median absolute difference (MADs) in group percentile ranks. This can be loosely interpreted as the typical difference that

switching between metrics would make in terms of percentile ranks. For example, the right panel of Figure 4a shows that switching between medSGPs and meanSGPs would change a typical group ranking by 4 percentile ranks using State B Reading data. The maximum data point in this boxplot also shows that, at worst in this dataset, switching between medSGPs and meanSGPs changes a group ranking by about 30 percentile ranks. In contrast, Figure 4b shows that for both simulated and empirical data, switching between FEMs and medPRRs or medSGPs results in a typical group ranking change of 6 percentile ranks with an interquartile range of about 2 to 11 percentile ranks and maximum change of about 40 percentile ranks. For example, for the State B reading data, one group's FEM is at or above about 82 percent of all the groups, making it appear as exceptional, whereas its medSGP places it at the 49th percentile of all the groups, a more mediocre ranking. Such schools could receive substantially different accolades or sanctions depending on which metric was used.

An alternative approach to describing metric dissimilarity is to describe the proportion of groups that are in the same quartile or decile by each pair of metrics, or the proportion of groups 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 group with a cross-metric change of 20 percentile ranks may or may not cross a quartile boundary. Figure 4 is a more robust representation that describes the magnitudes of typical and extreme changes in group percentile ranks.

Scale Invariance

To evaluate scale invariance, we use a family of four piecewise transformations as given in Castellano and Ho (2013). We apply these transformations directly to each of the standardized test score variables in a way that increases or decreases skewness $S(z)$ and kurtosis $K(z)$:

$$S(z) = \begin{Bmatrix} (z/k) + (1-k)/[k \times (k+1)]; & z < -1 \\ (2 \times z)/(k+1); & -1 \leq z < 0 \\ [(k+1) \times z]/2; & 0 \leq z < 1 \\ (k \times z) + (1-k)/2; & z \geq 1 \end{Bmatrix} \quad (6)$$

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

The constant *k* in these equations controls changes to the skewness and kurtosis of variables. 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. These choices mimic realistic values of skewness and kurtosis observed in practice, and these piecewise transformations follow those sometimes used by state testing programs (e.g. MDESE, 2009). 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. We use $J = 3$ prior years for this analysis to reflect a common number of prior years used in practice.

The 8 ACSMs are estimated for all transformed datasets. For a given dataset, each group has five values for each ACSM—one each from the original (identity), positive skew, negative skew, positive kurtosis, and negative kurtosis transformations. Each group is then rank ordered by each ACSM for each transformation, resulting in each group receiving five percentile ranks—one per transformation—for each ACSM. The range of these values indicates the extent that a group's percentile rank would change under this family of plausible transformations. We avoid correlations because they are limited to pairwise comparisons of transformations, and we are interested in a single metric describing the relevant variability of ranks under many plausible

transformations. Thus, we quantify the amount of transformation-induced variability by using mean ranges on a common scale of percentile ranks.

Figure 5 displays the average range of percentile ranks for each ACSM under the five transformations for each of the four empirical datasets. The most scale-invariant metric is the meanSGP metric. The medSGPs are comparable to the other metrics. This may seem counterintuitive, particularly because medians are generally recognized as being more stable under transformations. 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. This is akin to sampling variability, except there is no resampling. It is variance due to transformations, where rescaling leads to variance that looks similar to what we might expect from resampling. Although it may seem contradictory, at the aggregate level, the scale-dependent mean will maximize the scale-independent tendencies of SGPs.

Figure 5 also shows that meanPRRs are more robust to transformations than the mean-based residual metrics (FEM, REM, meanResid). This is due to the percentile rank transformation. When increases to skewness and kurtosis lead to extreme residuals, the PRR transformation will constrain the transformed residual within .1 and 99.9 no matter how large the residual is. Thus, averages of percentile ranks of residuals will be more stable under transformations than averages of residuals themselves. The medPRR metric performs relatively poorly for the same reason that medSGPs underperform meanSGPs. Note that the scale invariance of meanResids, FEMs, and REMs are as indistinguishable as the metrics themselves. With regard to comparisons across states, State A has larger mean ranges than State B, due in

part to having fewer groups (and thereby larger changes in percentile ranks) and increasing variance across grades.

Concluding Remarks

Although the literature on growth and value-added models is rich, focused 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 can support inferences about the status of a group by referencing current performance to expectations given prior scores. Empirical differences can thus be explained in terms of the approaches that different ACSMs take to support conditional status interpretations.

Our analyses support three empirically distinguishable categories of ACSMs. These categories are (1) the mean-based residual metrics (FEMs, REMs, and meanResids), (2) the mean-based percentile rank metrics (meanSGPs and meanPRRs), and (3) the median-based percentile rank metrics (medSGPs and medPRRs). The categories are listed in order of decreasing recovery of expected values (Figure 2). If we had included the medResid metric, it would belong in the third category due to its systematic similarity with medPRRs. If we had included the RAR metric, it would be in its own, fourth category, due to its predictable deviation from all other metrics in a context where within-group variability is substantial. All analyses reveal considerable similarities among ACSMs within each of the categories that we have identified and considerable dissimilarities between categories, particularly the category of median-based percentile rank metrics.

This paper is constrained in at least two important ways: it only considers prior test score variables as covariates and it only considers linear, homoscedastic mean regression models First, we restricted our focus to prior scores as covariates, to be in line with the operational estimation

of SGPs. However, the inclusion of other covariates such as student demographic covariates would change expected current status and may change the extent that our ACSMs of interest produce comparable results. Moreover, we use error-contaminated prior scores with no adjustment for measurement error in either the prior or current scores. This may bias each ACSM differentially and is an active area of current research (e.g., Akram, Erickson, & Meyer, 2013; Lockwood & McCaffrey, 2014).

Second, as Castellano and Ho (2013a) show, the relationship between SGPs and linear, mean-based ACSMs will decline as the assumptions of the latter are violated. If no accommodations to the model are made, 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. As Castellano and Ho (2013a) 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. Similarly, extending these analyses to incorporate student- and group-level covariates would be theoretically useful, even if many current state and federal policies discourage the resulting dependence of expectations upon covariates.

A consistent finding is that the aggregation function is a more consequential decision than the regression function or the choice between fixed and random intercepts. 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 three findings that suggest considerable advantages of means over medians. The first is the relative efficiency of the mean over the median (Figure 3). The second is the

greater alignment between meanSGPs/meanPRRs and their respective expected values (Figure 2) as well as their greater comparability with the FEM, REM, and meanResid metrics (Table 3 and Figure 4). The third is that meanSGPs and meanPRRs are more robust to scale transformations than their median-based counterparts (Figure 5).

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, we recommend a simple alternative that assumes that a normal distribution underlies the percentile rank function: $\widetilde{\text{meanSGP}} = \Phi\left(\sum \Phi^{-1}\left(\frac{\text{SGP}_i}{100}\right)/n\right)$, where Φ 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, and we have adhered to the simpler specification.

Descriptions of ACSMs often incorporate the terms “growth” and “value added.” These terms are ambiguous, and we recommend reviews such as those by Castellano and Ho for “growth” (2013b) and Reardon and Raudenbush for “value added” (2009) that specify assumptions and articulate the necessary data and model features to address specific questions. 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 the aggregation function is consequential. As stakes rise for accountability and evaluation decisions for teachers, schools, subgroups, and districts,

clear descriptions of ACSMs and the practical magnitudes of their dependencies become essential.

References

Akram, K., Erickson, F., & Meyer, R., (2013). Issues in the estimation of student growth percentiles. Paper presented at the annual meeting of The Association for Education Finance and Policy, New Orleans, LA. Retrieved from the AEFP website: http://www.aefpweb.org/

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessments of teachers. *Journal of Educational and Behavioral Statistics, 29,* 37– 65. doi: 10.3102/10769986029001037

Betebenner, D. W. (2008a). Toward a normative understanding of student growth. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

Betebenner, D. W. (2008b). *A primer on student growth percentiles.* Retrieved from the Georgia Department of Education website: http://www.doe.k12.ga.us/

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

Betebenner, D. W. (2010a). *New directions for student growth models*. Retrieved from the Kansas Department of Education website: http://www.ksde.org/

Betebenner, D. W. (2010b). *SGP: Student growth percentile and percentile growth projection/trajectory functions*. R package version 0.0-6.

Briggs, D., & Betebenner, D. W. (2009, April). *Is growth in student achievement scale dependent?* Paper presented at the invited symposium Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Casella, G., & Berger, R. L. (2001). *Statistical inference (2nd ed.)*. Pacific Grove, CA: Duxbury Thomson Learning.

Castellano, K. E., & Ho, A. D. (2013a). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi: 10.3102/1076998611435413

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

Colorado Department of Education. (2012). *Colorado growth model FAQs (General).* Retrieved from http://www.schoolview.org/GMFAQ.asp

Dette, H., & Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society B, 70(3),* 609 – 627. doi: 10.1111/j.1467-9868.2008.00651.x

Ehlert, M., Koedel, C., Parsons, E., & Podgursky. (2012). *Selecting growth measures for school and teacher evaluations.* Working Paper 80. National Center for Analysis of Longitudinal Data in Educational Research. Retrieved from the ERIC website: http://eric.ed.gov/?id=ED535515

Ellis, K. J., Abrams, S. A., & Wong, W. W. (1999). Monitoring childhood obesity: Assessment of the weight/height^2 index. *American Journal of Epidemiology, 150*, 939-946.

Fetler, M. E. (1991). A method for the construction of differentiated school norms. *Applied Measurement in Education, 4*, 53 – 66.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* Working Paper #18. The Education Policy Center at Michigan State University. Retrieved from

http://education.msu.edu/epc/library/documents/Guarino-Reckase-Wooldridge-May-2012-Can-Value-Added-Measures-of-Teacher-Performace-Be-Truste.pdf.

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2014). A comparison of growth percentile and value-added models of teacher performance. Working Paper #39. The Education Policy Center at Michigan State University. Retrieved from http://education.msu.edu/epc/publications/documents/WP39AComparisonofGrowthPercentileandValue-AddedModel.pdf.

Goldhaber, D., Walch, J., & Gabele, B., (2012). *Does the model matter? Exploring the relationship between different student achievement-based teacher assessments*. CEDR Working Paper 2012-6. University of Washington, Seattle, WA.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60 – 87. doi: 10.3102/0162373707299706

Houng, B., & Justman, M. (2013). *Comparing lest-squares value-added analysis and student growth percentile analysis for evaluating student progress and estimating school effects.* Melbourne Institute Working Paper Series Working Paper No. 7/13. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2230187

Kim, J.-S. and Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika*, *71*(4), 659–690.

Koenker, R. (2005). *Quantile regression*. New York, NY: Cambridge University Press.

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments*. Retrieved from https://www.parcconline.org/sites/parcc/files/PARCCVertScal289-12-201129.pdf.

Lissitz, R. W., Doran, H., Schafer, W. D., & Willhoft, J. (2006). Growth modeling, value added modeling and linking: An introduction. In R. W. Lissitz (Ed.) *Longitudinal and value added models of student performance* (pp. 1-46). Maple Grove, MN: JAM Press.

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39,* 22-52.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16,* 421-437. doi: 10.1177/001316445601600401

Manning, W. H., & DuBois, P. H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills, 15,* 287-321.

Massachusetts Department of Elementary and Secondary Education [MDESE]. (1999). *Massachusetts comprehensive assessment system: 1998 technical report*. Retrieved from http://www.doe.mass.edu/mcas/tech/98techrpt.pdf

Massachusetts Department of Elementary and Secondary Education [MDESE]. (2009). *MCAS student growth percentiles: State report*. Retrieved from http://www.doe.mass.edu/mcas/growth/StateReport.pdf

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

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606. doi: 10.1162/edfp.2009.4.4.572

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48,* 163-193. doi: 10.3102/0002831210362589

Rabe-Hesketh, S., & Skrondal, A. (Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata (3rd Edition): Volume I: Continuous responses*. College Station, TX: Stata Press.

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

R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *American Education Finance Association, 4*(4), 492 – 519. doi: 10.1162/edfp.2009.4.4.492

Sanders, W. L., Saxton, A. M., & 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: Corwin Press.

Sanders, W. L. (2006). *Comparisons among various educational assessment value-added models.* (SAS White Paper). Retrieved from http://www.sas.com/resources/

Scholten, A. Z., & Borsboom, D. (2009). A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology, 53*, 69-75. doi: 10.1016/j.jmp.2009.01.002

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Washington, D.C.: Sage Publications, Inc.

Spencer, B. D. (1983). Test scores as social statistics: Comparing distributions. *Journal of Educational Statistics, 8*(4), 249-269.

Suen, W. (1997). Decomposing wage residuals: Unmeasured skill or statistical artifact? *Journal of Labor Economics, 15(3),* 555 – 566. doi: 10.1086/209872

United States Department of Education. (2005). Secretary Spellings announces growth model pilot, Addresses chief state school officers' annual policy forum in Richmond (press release). Retrieved from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html

United States Department of Education. (2009). *Race to the top program: Executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* Cambridge, MA: The MIT Press.

Wright, S. P., (2010). An investigation of two nonparametric regression models for value-added assessment in education. SAS White Paper. Retrieved from http://www.sas.com/whitepapers/indexAZ.html

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 205 – 218.

Table 1

Table of Aggregate-Level Conditional Status Metrics

Abbreviation	Metric	Regression Approach	Aggregation Function	Scale for Interpretation
medSGP	Median Student Growth Percentile	Nonlinear Quantile	Median	Percentile Rank
meanSGP	Mean Student Growth Percentile	Nonlinear Quantile	Mean	Percentile Rank
medPRR	Median Percentile Rank of a Residual	Linear Mean	Median	Percentile Rank
meanPRR	Mean Percentile Rank of a Residual	Linear Mean	Mean	Percentile Rank
meanResid	Mean Residual	Linear Mean	Mean	"Current" Score Scale (Residual)
medResid	Median Residual	Linear Mean	Median	"Current" Score Scale (Residual)
RAR	Residuals from Aggregate Regression	Linear Mean	Mean	"Current" Score Scale (Residual)
FEM	Fixed Effects Metric	Linear Mean (incorporates group membership)	Mean	"Current" Score Scale (Residual)
REM	Random Effects Metric	Linear Mean (incorporates group membership)	Mean	"Current" Score Scale (Residual)

Note: In this paper, for all regression approaches, current status, or the current grade-level score, is the response variable and prior grade-level scores are the predictors.

Table 2

Summary of the Recovery of each Aggregate-Level Conditional Status Metric's Parameter as measured by Root Mean Square Error (RMSE) and Pearson Correlations.

			Percentile-Rank Metrics				Residual-based Metrics		
Out-come	N_g	ICC	Med-SGP	Med-PRR	Mean-SGP	Mean-PRR	MeanResid (in σ_Y units)	FEM (in σ_Y units)	FEM and meanResid (in $\mathrm{SD}(\mu_{\epsilon\|g})$ units)
RMSE	25	0.05	9.36	9.36	5.68	5.69	0.111	0.111	1.10
		0.15	8.83	8.82	5.50	5.50	0.107	0.107	0.61
		0.25	8.28	8.27	5.31	5.31	0.103	0.103	0.46
	50	0.05	6.66	6.67	4.02	4.02	0.078	0.078	0.78
		0.15	6.28	6.28	3.91	3.91	0.076	0.076	0.44
		0.25	5.95	5.95	3.81	3.8	0.074	0.074	0.33
	100	0.05	4.76	4.76	2.83	2.83	0.055	0.055	0.55
		0.15	4.52	4.52	2.78	2.78	0.054	0.054	0.31
		0.25	4.29	4.29	2.71	2.71	0.053	0.053	0.23
Correlation	25	0.05	0.59	0.59	0.66	0.66	0.67	0.67	
		0.15	0.79	0.79	0.85	0.85	0.85	0.86	
		0.25	0.87	0.87	0.91	0.91	0.91	0.91	
	50	0.05	0.72	0.72	0.78	0.78	0.79	0.79	
		0.15	0.88	0.88	0.91	0.91	0.92	0.92	
		0.25	0.93	0.93	0.95	0.95	0.95	0.95	
	100	0.05	0.83	0.83	0.87	0.87	0.88	0.88	
		0.15	0.93	0.93	0.95	0.95	0.96	0.96	
		0.25	0.96	0.96	0.97	0.97	0.97	0.98	

Note. All of the metrics were unbiased; that is, their bias was not significantly different from 0. The RMSEs for mean and median SGP and PRRs are on the percentile rank scale. The RMSEs for meanResids and FEMs are expressed in terms of standard deviation units of individual scores, σ_Y, or $\mathrm{meanResid}_g^{\mathrm{true}}$, $\mathrm{SD}(\mu_{\epsilon|g})$, as indicated.

Table 3

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

Metric	REM	FEM	Mean Resid	Mean PRR	Mean SGP	Med PRR	Med SGP
Simulated MVN Data – Group Sizes = 25 (ICC=.05 above/ICC=.25 below)							
REM		1.000	1.000	0.986	0.985	0.868	0.868
FEM	1.000		1.000	0.986	0.985	0.868	0.867
meanResid	0.999	0.998		0.986	0.985	0.868	0.868
meanPRR	0.994	0.994	0.996		0.999	0.909	0.908
meanSGP	0.994	0.993	0.995	0.999		0.908	0.908
medPRR	0.950	0.949	0.951	0.964	0.963		0.991
medSGP	0.949	0.949	0.951	0.963	0.964	0.996	
State A Data (Reading above/Mathematics below)							
REM		0.993	0.993	0.980	0.977	0.892	0.900
FEM	0.996		1.000	0.988	0.986	0.896	0.903
meanResid	0.996	1.000		0.988	0.986	0.897	0.903
meanPRR	0.988	0.992	0.992		0.997	0.935	0.938
meanSGP	0.987	0.991	0.991	0.999		0.932	0.941
medPRR	0.944	0.948	0.948	0.965	0.965		0.975
medSGP	0.943	0.948	0.947	0.964	0.965	0.994	
State B Data (Reading above/Mathematics below)							
REM		0.995	0.995	0.985	0.985	0.923	0.930
FEM	0.999		0.999	0.989	0.989	0.926	0.933
meanResid	0.998	0.999		0.991	0.989	0.928	0.934
meanPRR	0.995	0.995	0.997		0.997	0.954	0.956
meanSGP	0.992	0.993	0.995	0.998		0.953	0.959
medPRR	0.975	0.975	0.977	0.984	0.981		0.989
medSGP	0.976	0.976	0.978	0.984	0.986	0.990	

Note. meanSGP = mean Student Growth Percentile; medSGP = median Student Growth Percentile; meanPRR = mean Percentile Rank of Residual; medPRR = median Percentile Rank of Residual; meanResid= Group-mean Residuals from linear mean regression; medResid = Group-median residuals from linear mean regression; FEM = Fixed-Effect Metric; REM = Random-Effect Metric.

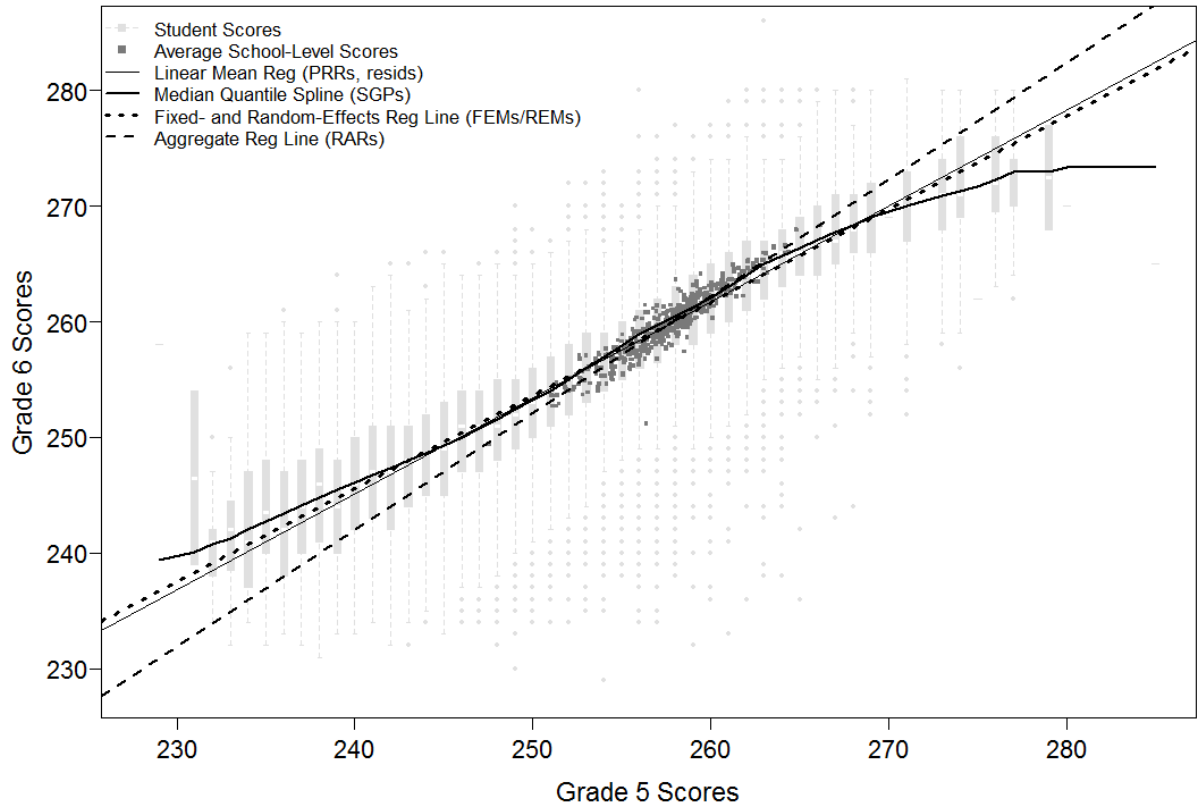

Figure 1. Contrasting the models used in deriving aggregate-level conditional status metrics for a $J = 1$ prior-grade empirical test score dataset. The student scores are expressed as conditional boxplots of current score on initial status, and the school mean scores are overlaid as solid grey squares.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

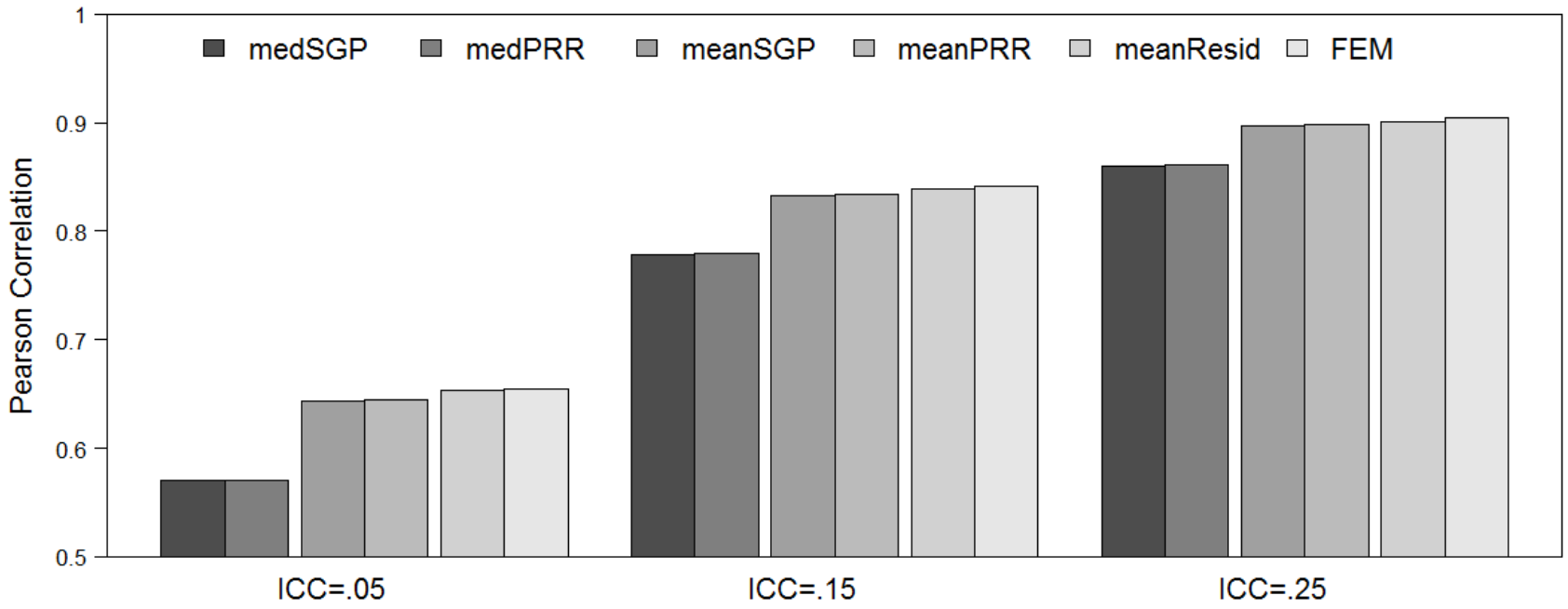

Figure 2. Pearson correlations between each Aggregate-Level Conditional Status Metric and its respective parameter value, averaged over 100 replications of simulated multivariate normal data with 500 groups of sizes 25, 50, and 100 and unconditional intraclass correlations (ICCs) of ω =.05, .15, and .25.

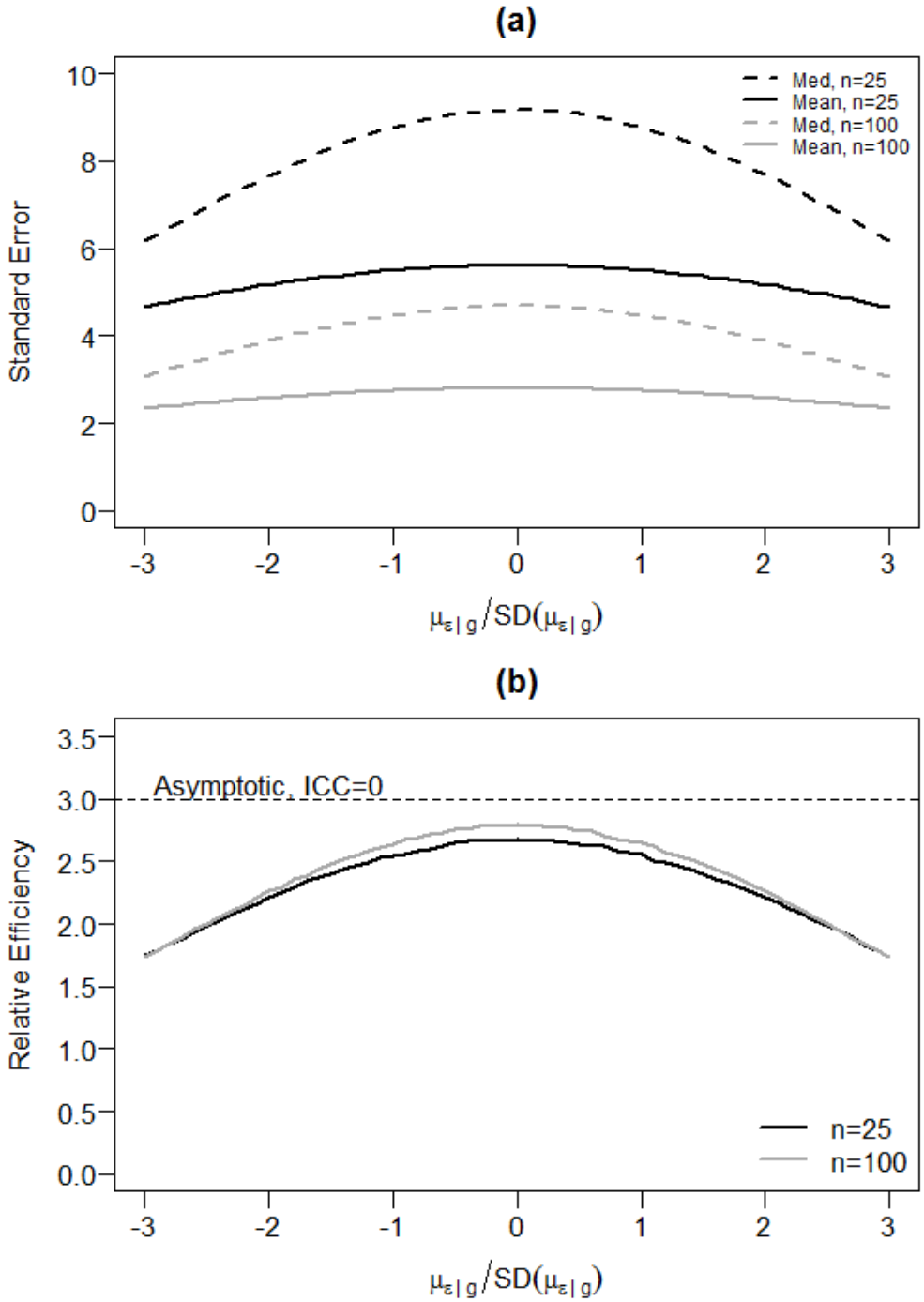

Figure 3. Standard error (a) and relative efficiency (b) of mean vs. median percentile ranks of residuals when residual group means are relatively low or high in terms of standard deviation units of the distribution of $\text{meanResid}_g^{\text{true}}$, $SD(\mu_{\epsilon|g})$. Intraclass correlation (ICC) of $\omega = .15$. Asymptotic relative efficiency is 3 when $\omega = 0$ and is shown for reference.

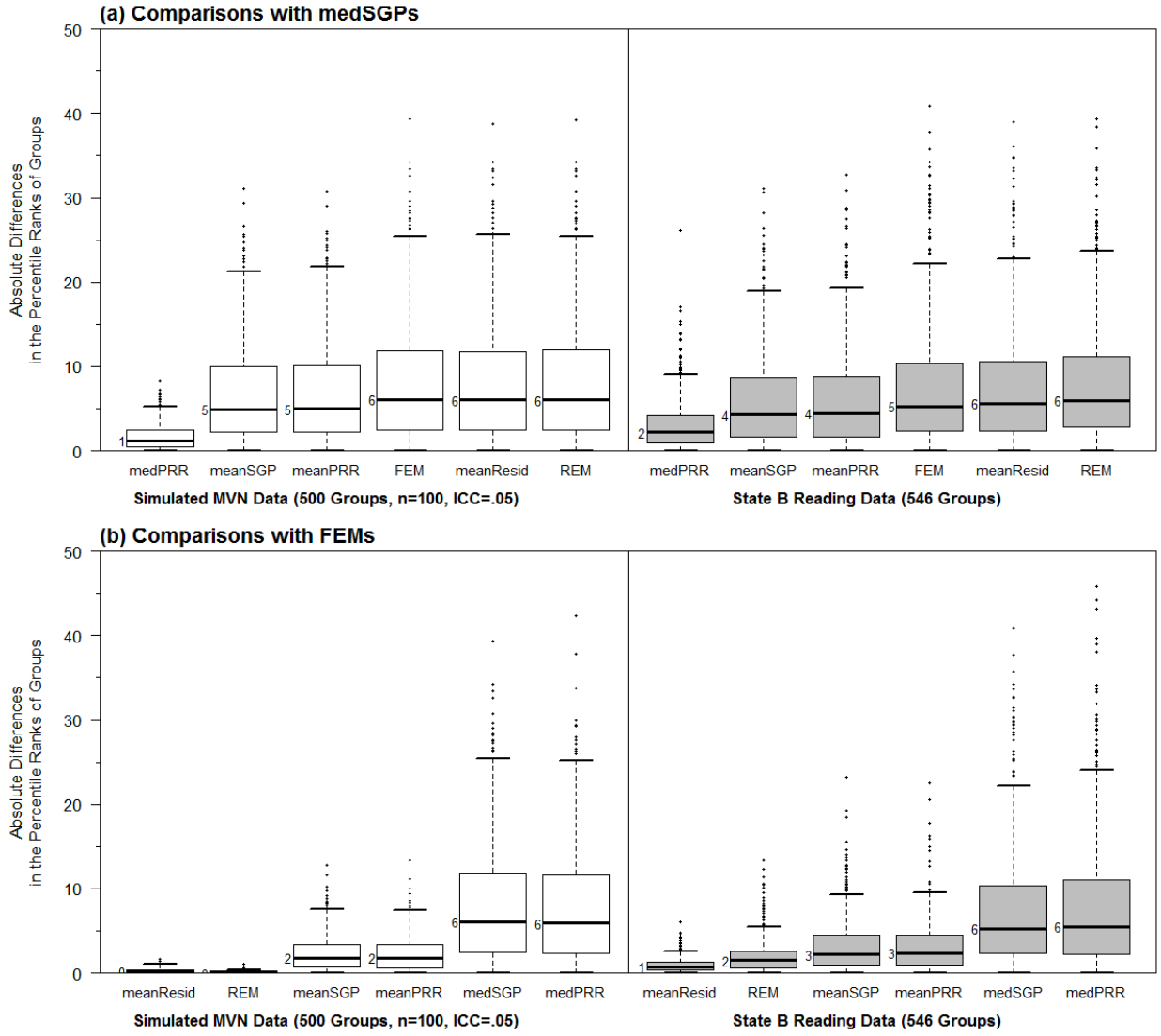

Figure 4. Comparing each metric with (a) the median Student Growth Percentile (medSGP) metric and (b) the fixed effect metric (FEM) in terms of absolute differences in the percentile ranks of groups. Findings for simulated data are on the left (for a single replication of simulated multivariate normal data, ICC=.05, 500 groups of size 100), and findings for State B Reading data are on the right. medPRR=median Percentile Rank of Residuals; meanSGP=mean Student Growth Percentile; meanPRR=mean Percentile Rank of Residuals; REM=Random-Effects Metric; meanResid=mean residuals.

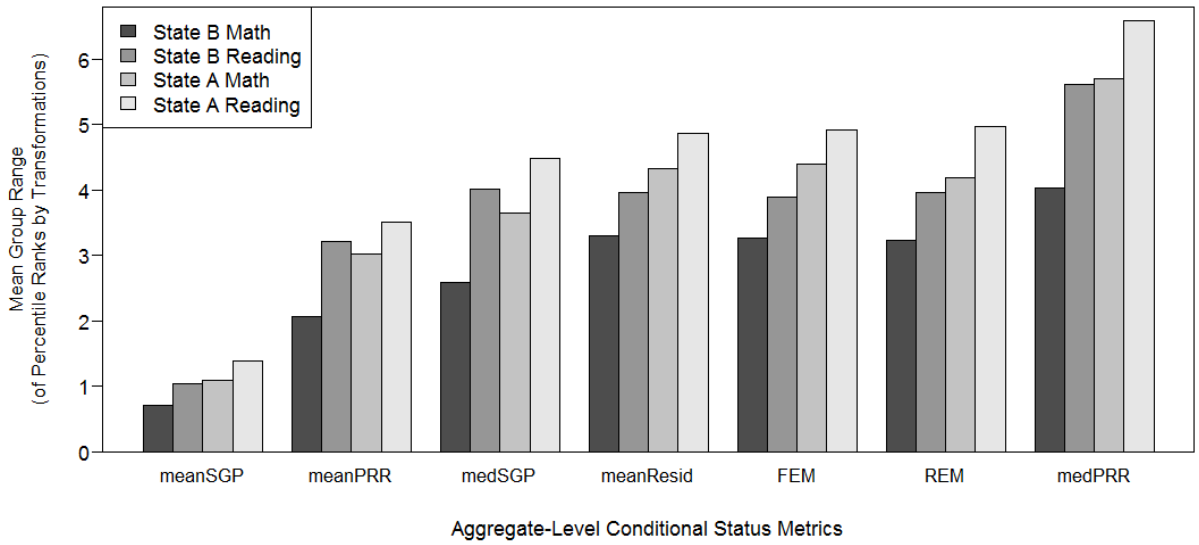

Figure 5. The range of percentile ranks for an average group, under a family of plausible transformations that change skewness and kurtosis. From left, meanSGP = mean Student Growth Percentile, meanPRR = mean Percentile Rank of Residual, medSGP = median Student Growth Percentile, meanResid = mean Residuals, FEM = Fixed-Effects Metric, REM = Random-Effects Metric, and medPRR = median Percentile Rank of Residuals.