Sensitivity of School-Performance Ratings to Scaling Decisions

Sensitivity of School-Performance Ratings to Scaling Decisions
(With Supplementary Materials)

Hui Leng Ng[1]

Harvard Graduate School of Education

Phone: (+65) 9664-2439

Fax: -

Email: hui_leng_ng@mail.harvard.edu

Daniel Koretz

Harvard Graduate School of Education

415 Gutman Library

6 Appian Way

Cambridge, MA 02138

Phone: (617) 384-8090

Fax: (617) 496-3095

Email: daniel_koretz@gse.harvard.edu

Suggested running head: Sensitivity of school ratings to scaling

Acknowledgement

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305AII0420, and by the Spencer Foundation, through Grants 201100075 and 201200071, to the President and Fellows of Harvard College. The authors also thank the New York State Education Department for providing the data used in this study. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, the Spencer Foundation, or the New York State Education Department or its staff.

---

[1] Hui Leng Ng is currently at the Singapore Ministry of Education. She may be contacted at hui_leng_ng@mail.harvard.edu or 285 Ghim Moh Road, Singapore 279622. The opinions expressed in the article are those of the authors and do not represent the views of the Singapore Ministry of Education.

Sensitivity of School-Performance Ratings to Scaling Decisions

Abstract

Policymakers usually leave decisions about scaling the scores used for accountability to their appointed technical advisory committees and the testing contractors. However, scaling decisions can have an appreciable impact on school ratings (Briggs & Weeks, 2009). Using middle-school data from New York State, we examined the consistency of school ratings based on two scaling approaches that differed in scaling decisions that are important in high-stakes testing contexts. We found that, depending on subject, grade, and year, a switch in scaling approach led to (1) average absolute shifts in ranks of between 50 and 132 positions (median = 69), which are appreciable shifts for a listing of 1,243 schools; and (2) between 7% and 45% (average = 20%) of schools experiencing shifts in assigned performance bands, depending on the classification scheme. Further, the effect of scaling approach was larger when the raw-score distribution has more severe ceiling effect, and in these cases, it was driven primarily by the difference in the location of the highest obtainable scale score from the two scaling approaches.

Sensitivity of School-Performance Ratings to Scaling Decisions

Introduction

Policymakers who commission standardized tests usually leave the choice of the scaling approach used to generate the scores to their appointed technical advisory committee and the testing contractor (National Research Council, 2010). Given the heavy reliance on the resulting scores for inferences about educators' relative performance, we argue that it is important to examine the robustness of such inferences to reasonable alternative scaling decisions because such decisions are typically substantively unrelated to the target inferences.

Past studies have investigated the differences in the scale scores obtained from different ability estimators (e.g., Kim & Nicewander, 1993). But these studies did not examine the impact of such differences in the resultant scale scores on school-performance ratings.

A few recent studies have investigated the sensitivity of teacher or school value-added measures to the psychometric properties of the underlying scale. But these studies tended to focus on issues related to the violations of interval-scale or vertical-scale properties (Ballou, 2009; Briggs & Betebenner, 2009; Martineau, 2006), rather than to reasonable alternative scaling decisions made during the scaling process.

One notable exception is a study by Briggs and Weeks (2009), who examined the sensitivity of schools' value-added estimates to decisions about the scaling model, linking method, and the estimation method when creating a vertical scale. The resulting estimates were strongly linearly inter-related (Pearson correlations between .79 and .99) but nonetheless often resulted in appreciably different classifications of schools into three broad performance bands.

This study, like that of Briggs and Weeks (2009), focuses on the impact of scaling decisions, but it addresses two issues that commonly arise in practice in high-stakes testing:

(1) right-censoring of the raw-score distribution (i.e., a ceiling effect); and (2) the use of methods that create a 1-to-1 mapping between raw and scale scores.

Ceiling effects are frequently observed in high-stakes testing programs (Ho & Yu, 2012). They arise because of the initial easiness of the test, the typically rapid rise in scores (often a result of score inflation), or both. For example, Koedel and Betts (2010) reported that, in 2006, the high-school exit examinations in 26 states were "pitched at a middle school or lower high school level" (p. 55). Similarly, many researchers have documented performance gains on high-stakes state tests that far outpaced the gains on a lower-stakes test (e.g., NAEP) over the same time period, suggesting the presence of score inflation (e.g., Fuller, Gesicki, Kang, & Wright, 2006; Jacob, 2007). These rapid gains substantially exacerbate ceiling effects.

Ceiling effects on the raw-score distribution lead to unavoidable uncertainties in scaling raw scores near the ceiling. This uncertainty is exacerbated when the choice of a scaling method requires setting the lowest and highest obtainable scale score (LOSS and HOSS) *a priori*. While the transformation of raw scores into scale scores by IRT methods mitigates the ceiling effects by stretching out the upper tail of the distribution, different scaling methods are likely to stretch the tail by varying amounts, and it is unclear which amount is the most reasonable.

To avoid confusion and to enable practitioners to convert between raw scores and scale scores, a majority of states adopt methods that produce a 1-to-1 mapping between raw and scale scores. Many use Rasch scoring, which directly produces such a mapping because raw scores are a sufficient statistic for the Rasch ability estimates. However, some states that use IRT pattern

scaling nonetheless also report scores, commonly called summed scores, with a 1-to-1 mapping to raw scores. In these cases, an extra step is required to convert the θ scale to summed scores. [1]

In this study, we used data from the New York State testing program to investigate whether the process of obtaining summed scores coupled with pattern scaling is of practical importance to schools' ratings, and we explored the interaction of this impact with the severity of the ceiling effects on the raw-score distributions. We examined the consistency of school ratings based on two different scaling approaches. One approach, used operationally by the state, employed maximum-likelihood estimation and therefore set the LOSS and HOSS *a priori*. Summed scores were then created by inverting the test-characteristic curve obtained from pattern scaling (see later). This resulted in large differences in summed scores at both ends of the distribution between students whose raw scores differed by only a single raw-score point. The second approach produced scale scores that had neither a 1-to-1 mapping nor large gaps between scale scores at the extremes.

Besides focusing on scaling decisions that are particularly important in high-stakes testing contexts, our study also differed from the study by Briggs and Weeks (2009) in three other respects. First, instead of the multivariate "layered model" for value-added measures that the authors employed, we used covariate-adjustment measures.[2] Second, instead of reading

---

[1] Of the 36 states for which we found the relevant technical information spanning school years 2006-07 to 2010-11, 30 reported scales with a 1-to-1 mapping to raw scores. Of these, 18 used the Rasch model, while the others used pattern scaling coupled with a process of obtaining a 1-to-1 mapping to raw scores. Details are in Appendix A of the supplementary materials available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:13360004.

[2] We conducted similar analyses using difference-score measures, that is, using the difference between current- and prior-year scores as the dependent variable. The results were largely similar and are available upon request.

achievement, we used both English language arts (ELA) and mathematics. Finally, in addition to the school-classification scheme that they used, which we report in detail below, we also used two other classification schemes, chosen because they have been used by other researchers or in existing school-accountability systems.

In these analyses, we addressed three specific research questions:

RQ1. *How do the two scales differ in addressing the ceiling effects on the raw-score distributions?*

RQ2. *How consistent are schools' performance ratings in a specific subject, grade, and year, when we use scores from the two different scales to derive the school-performance estimates?*

RQ3. *What is the relative importance of the various differences in scaling decisions made while creating the two scales in contributing to the inconsistency in schools' ratings?*

Methodology

Data

The NYS dataset that we used contained student-level ELA and mathematics performance and student demographics, for all students participating in the NYS accountability-testing program in grades 7 or 8 in school years ending Spring 2009 and 2010. Performance data included both item responses and scale scores provided by NYS's testing contractor and used operationally in NYS (henceforth, the "TC scale"). This scale was set to a fixed mean and standard deviation in every grade and subject in the first year of the program, and the data in the subsequent years were linked back to this initial scale. NYS uses this scale to implement its

accountability system. We rescaled the performance data to create a second set of scale scores (henceforth, the "Alt scale") solely for research purposes.[3] In every grade and subject, this scale was initially approximately mean zero and standard deviation one in the first year for which we had data, and was linked across years using the same linking constants as those used to create the TC scale.

The Two Scaling Approaches

Both the TC and Alt scales were derived from three-parameter logistic IRT models for the multiple-choice items. This is a pattern-scoring IRT model that specifies the probability of getting an item correct as a logistic function of the test-taker's proficiency (i.e., the latent ability, θ), and three parameters describing the difficulty, discrimination, and pseudo-guessing rate of the item. Further, the same linking constants were used to link each of the two scales over time.

However, the two approaches differed in (1) the model used for constructed-response items; (2) the approach used to estimate the item parameters (i.e., item calibration); and (3) the approach used to obtain students' scale scores. The third is the most important difference.

For the constructed-response items, NYS's testing contractor used a two-parameter generalized partial-credit (GPC) model, while we used a graded-response model, selected for reasons unrelated to this paper. The GPC model estimates the probability of a response falling into any single score category. The graded-response model estimates the probability of scoring at a given step or higher; probabilities for individual steps can then be obtained by subtraction. However, past research has shown that the two types of models produce very similar scores (Maydeu-Olivares, Drasgow, & Mead, 1994; Thissen, Nelson, Rosa, & McLeod, 2001).

[3] We used IRTPro, developed by Li Cai, David Thissen and Stephen du Toit, to derive the Alt and Alt1-1 (see later) scale scores.

Secondly, the two scales differed in their item-calibration approaches. NYS's testing contractor used marginal maximum-likelihood estimation, while we used a Bayesian approach that allows the specification of a prior distribution for each of the item parameters. However, perhaps because of the large amount of data, we found that priors for the item discrimination and item difficulty (*a*- and *b*-parameters) had no appreciable effect, so we did not specify them, effectively setting uniform priors for those parameters. We only specified a Beta(6,16) prior distribution for the *c*-parameter. We would therefore expect this difference between the two scales to play a minor role in driving any observed scaling-approach effect on schools' ratings, and across all combinations of grade, subject, and year used in the study, the median correlation between estimates of the *a*- and *b*-parameters obtained for multiple-choice items for the two scales were both .97.[4]

Finally, the two scales differed in two aspects of the methods used to create students' scale scores. First, maximum likelihood methods were used to derive the TC scale scores, while we used expected *a posteriori* (EAP) estimation to derive the Alt scale scores. Second, the TC scale scores are not simple linear transformations of θ. Rather, being summed scores, they are discrete estimates of θ. In contrast, the Alt scale scores are simple linear transformations of θ.

Because the TC scale scores were estimated using maximum-likelihood methods, θ is undefined for students whose raw scores are either zero or perfect, and the latter are numerous when raw scores show a ceiling effect. Scale scores for these students—the LOSS and HOSS—therefore must be set arbitrarily (CTB/McGraw-Hill, 2006). In addition, LOSS was also assigned

[4] Although the corresponding correlations between estimates of the *c*-parameter were lower (median = .46), these are likely to be attenuated due to the greater uncertainty associated with estimations of the *c*-parameter, compared to the *a*- or *b*-parameters, in general, as documented by past research (Han, 2012; McKinley & Reckase, 1980).

to students scoring below chance. The LOSS and HOSS were kept constant across years. In contrast, Bayesian approaches, such as that used to obtain our Alt scale scores, estimate θ directly for zero and perfect scorers.

The TC summed scores were obtained by inverting the test characteristic curve (TCC). This process begins with estimation of the item characteristics curves (ICCs) that show the estimated performance on each item as a function of θ (for binary items, the probability of a correct response). The TCC is the sum of the ICCs across all items. This provides at any value of θ the "number right true score" (NRTS)—the expected raw score for students with that value of θ (Lord, 1980). However, the NRTS, unlike a raw score, is not limited to integer values.

The approach used to obtain the TC summed scores works backwards from the NRTS. Each observed (integer) raw score is mapped to the TCC. The value of θ that would produce an NRTS equal to that raw score is then assigned to all students with that observed raw score. This produces a set of summed scores that maps 1-to-1 with the observed raw scores. Finally, this discretized, estimated θ distribution is linearly transformed to the reporting scale. Students with perfect scores were assigned HOSS; those with zero scores or scores below chance were assigned LOSS.

In contrast, we used EAP estimation to obtain the Alt scale scores, setting priors for the ability distribution. We did not create summed scores for our primary Alt scale. However, to test the impact of discretizing the whole scale with 1-1 mapping to the observed raw scores, we also created summed scores using the EAP estimation method described by Thissen and Orlando (2001, pp. 119-121). This scale, which we label Alt1-1, is identical to the Alt scale except for the reduction to summed scores.

The resulting TC and Alt scale scores differ in two respects. First, while the TC scale scores have a 1-to-1 mapping to the raw scores for all students, the Alt scale scores have a 1-to-1 mapping only for zero or perfect raw scores. (This mapping occurs because all students with zero or perfect scores had the same response patterns, either all answers incorrect or all answers correct.) Second, the two sets of scale scores differed in the amount of stretching in the two tails of the distributions. This difference is in addition to the shrinkage inherent in Bayesian estimates such as the Alt scale; it is a non-uniform difference that becomes progressively more extreme as scores deviate further from the mean. Specifically, for the TC scale scores, the inversion of the TCC and the adoption of LOSS and HOSS resulted in increasing distances between the scale scores corresponding to two adjacent raw scores. This was true at both ends of the distribution but was more evident at the upper end because of the ceiling effect on raw scores. The Alt scale scores do not show such increasing or large gaps between two adjacent scale scores at either end of the distribution. In Figure 1, we illustrate these differences, using the 2009 ELA results of grade-7 students. The increasing and large gaps at both ends on the TC scale-score distribution in the sample—and their absence in the corresponding Alt scale-score distribution—are evident from both the histograms for the two sets of standardized scale scores (top panel) and the scatter plot of the standardized Alt scale scores versus the standardized TC scale scores (bottom panel).

~~~~~ INSERT **FIGURE 1** ABOUT HERE ~~~~~

Measures

In Appendix A, we display the principal variables that we used.

Outcome variables. The outcome variables were TC and Alt scale scores in ELA and mathematics on the state tests in the target year (Y = 2009 or 2010), separately for each subject. We denote each of these by $SCORE(Y)$.
~~~~~

Control predictors. The covariate adjustment models included various combinations of the following:

Achievement in Previous Year. TC and Alt scale scores in the same subject in the year before the target year. We denote this by $SCORE(Y-1)$.

Prior Achievement in Milestone Grade (G5). In some models, we included students' achievement in ELA and mathematics when they were in grade 5, which is typically the final grade prior to a student's entry to middle school. These covariates were on the same scale as the outcome variable in the model. We denote these covariates collectively by the vector $\boldsymbol{PA}$. Following the approach taken by Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007), $\boldsymbol{PA}$, unlike $SCORE(Y-1)$, included scores in *both* ELA and mathematics. $\boldsymbol{PA}$ served as a proxy measure of students' academic achievement at the end of elementary school.

Because the performance measures are on different scales, we standardized each of them with reference to the performance of a selected anchor cohort of students, separately for each combination of scale, subject, grade, and year. We used the 2006 grade-5 cohort as the anchor cohort, 2006 being the earliest year that we have access to in the NYS database, and grade 5 being the earliest grade whose data we used in the study. For example, a grade-7 student with a standardized TC scale score of 1 unit on the mathematics test in 2009 is 1 SD above the average score on the TC scale of the 2006 grade-5 cohort when the latter took the grade-7 mathematics state-test in 2008. This is akin to using a norming sample in the creation of a scale with normative interpretations (Kolen, 2006).

Student Background. Three sets of dichotomously coded covariates recorded selected student-background characteristics: (1) gender, family-income status, immigrant status, and several race/ethnicity categories; (2) limited-English-proficient (LEP), and disability statuses;

and (3) whether the student had received testing accommodations while taking the state tests. We denote these covariates collectively by the vector $\boldsymbol{B}$.

School-level Aggregate Variables. We also derived aggregate school-level measures by averaging the corresponding student-level variables within each school. We denote these collectively by the vector $\boldsymbol{S}$.

Construction of Analytic Sample

The total sample comprised 765,843 7th and 8th graders in 2009 and 2010 in 1,451 schools, roughly evenly distributed between the two grades and the two years. We constructed our analytic sample by first eliminating students with missing values for any variables needed to compute the school-performance estimates. Then we retained only schools that contained students in both grades who satisfied the student-level inclusion criteria for both subjects in both target years. This created a common set of schools for computing the school-performance estimates for all subjects, grades, and years. This is essential because normative school-performance measures depend on the particular schools included in the estimation sample.

The resulting analytic sample comprised 661,504 7th and 8th graders in 2009 and 2010 in 1,243 schools. This represents attrition rates of 14% at both the student and school levels. Nonetheless, the analytic sample is comparable to the total sample with regard to all student-background variables, at both the student and school levels.[5] For example, in terms of race/ethnicity, the total (analytic) sample comprised 54% (56%) White, 18% (17%) African-American, 20% (18%) Hispanic, and 8% (9%) Asian or others. Similarly, 48% (46%) of the students in the total (analytic) sample were from low-income families. But students in the

[5] Details, including school-level descriptive statistics, are available Appendix B of the supplementary materials available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:13360004.

analytic sample were slightly higher performing on average than those in the corresponding total sample, with difference in average scores ranging from .01 SD to .06 SD (average = .04 SD) across all performance measures.

Evaluating Ceiling Effects

The differences between the two scales in the upper tail of the distribution are of particular importance because of ceiling effects on the raw-score distributions. We quantified the severity of the ceiling effects three ways: the magnitudes of negative skewness (following Koedel and Betts, 2010), more positive kurtosis, and the distance from the median to the maximum standardized scores on the two sets of scale scores.

Creating School-Performance Measures

We defined a set of school-performance measures using covariate-adjustment models, with prior achievement as one of the covariates. We fitted six types of models defined by the control predictors included (Table 1). We focus here on the model that included all control predictors (CA6), but we used models that omitted one or more predictors to serve as sensitivity checks on model dependence.

~~~~~ INSERT **TABLE 1** ABOUT HERE ~~~~~

For each combination of scale, subject, grade, year, and set of covariates, we generated the school-performance estimates by fitting this 2-level random-intercepts multilevel model:

$$\textbf{(1)}\ SCORE(Y)_{is} = \mu + \alpha SCORE(Y-1)_{is} + \beta' B_{is} + \pi' PA_{is} + \gamma' S_s + \psi_s + \varepsilon_{is}$$

for student *i* in school *s*, where β', π', and γ' represent the vectors of coefficients for the student background variables, prior achievement variables, and school-level aggregate variables
~~~~~

respectively. All variables are grand-mean-centered.[6] For each combination of scale, subject, grade, year, and set of covariates:

- ε_{is}, the residual error term for student *i* in school *s*, is assumed to be independent and normally distributed with mean zero and variance σ_{ε}^2, for all *i* and *s*.
- ψ_s, the estimate of the performance of school *s*, is the empirical Bayes residual, i.e., the shrunken deviation of the school's mean performance from its performance predicted by the model specified in equation (1) (Raudenbush & Bryk, 2002). We assumed these school-performance estimates to be independent of ε_{is} for all *i*, and *s*, and that they were drawn from a normal distribution with mean zero and variance σ_{ψ}^2.

Estimating the Impact on Schools' Performance Ratings

We estimated the impact of the choice of scaling approach on two common uses of school-performance measures: (1) to create rank-ordered lists of schools; and (2) to classify schools into broad performance bands.

Impact on Schools' Ranks. We used two indices to quantify the impact of the choice of scaling approach on schools' ranks. First, we computed the Spearman's *rho* (rank correlations) between school-performance estimates obtained from the two scales:

$$\textbf{(2)}\ r_s = Corr\left(Rank(\hat{\psi}_s \mid_{\text{TC scale}}), Rank(\hat{\psi}_s \mid_{\text{Alt scale}})\right)$$

[6] Following the argument by Thum and Bryk (1997), we fitted only multilevel models with fixed coefficients for all predictors because we were only interested in a school's performance averaged over all its students.

where $Rank(\hat{\psi}_s \mid_{\text{TC scale}})$ and $Rank(\hat{\psi}_s \mid_{\text{Alt scale}})$ denote school ranks on the school-performance estimates derived from the TC scale scores and Alt scale scores respectively. Secondly, we computed the mean absolute difference between the two ranked variables:

$$\textbf{(3)}\ MAD = \frac{1}{N}\sum_{s=1}^{N}\left|Rank\left(\hat{\psi}_s \mid_{\text{TC scale}}\right) - Rank\left(\hat{\psi}_s \mid_{\text{Alt scale}}\right)\right|$$

This index represents the average shift in school ranks in either direction when one set of scale scores derived from one scaling approach was replaced with the other.

Impact on Schools' Assignment to Performance Bands. The impact of scaling approach on the assignment of schools to performance bands will depend on the classification scheme employed. In general, the proportion of schools changing classifications will be lower if there are fewer cut scores, if the cuts are in parts of the distribution with low density, or if the marginal distributions are substantially non-uniform. Therefore, specific classification schemes served only as illustrations of possible impact. We examined three schemes, selected for their uses in past research or school-accountability systems, but focused on two of them here. In Scheme 1, we classified schools with school-performance estimates that are at least one posterior SD (i.e., the SD of the distribution of empirical Bayes residuals obtained from equation [1]) below the average as "below average"; those with estimates that are at least one posterior SD above the average as "above average"; and all other schools as "average". This scheme was used by Briggs and Weeks (2009) and has been used frequently in effective-schools research to identify "outlier" schools (Crone, Lang, & Teddlie, 1995). Quantiles are often used in teacher/school value-added studies to illustrate the amount of classification inconsistency

associated with a correlation between alternative value-added measures (e.g., Ballou, 2009; Corcoran et al., 2011; Papay, 2011). For Scheme 2, we used quintiles. [7]

We computed the average percentage agreement between the TC and Alt scale scores for both schemes, separately for each combination of subject, grade, year, and model specification. We compared these percentages to the percentage of chance agreement (that is, the agreement rate expected with random assignment of schools to the performance bands), which is a function of the number of cut scores and the marginal distributions.

Investigating the Relative Importance of Different Scaling Decisions

We investigated the relative importance of three aspects of the scaling methods: (1) the reduction of the scale to summed scores; (2) inverting the TCC; and (3) setting the LOSS and HOSS *a priori*.

We evaluated the effects of reduction to summed scores by comparing the impact of the Alt and Alt-1 scores. Because the Alt1-1 scale is identical to the Alt scale except for the reduction to summed scores, if the effects of substituting the Alt1-1 scale scores for the TC scale scores are very similar to the effects of substituting with the Alt scale scores, then the scaling-approach effects must be attributable to some combination of the second and third factors.

---

[7] In addition, unequal proportion, asymmetric classification schemes are sometimes used in practice. For Scheme 3, we adapted the system used in in New York City's Progress Report for schools (New York City [NYC] Department of Education, 2011a, 2011b). For the 2010-11 school year, NYC's elementary and middle schools were assigned letter grades according to the following percentile ranks: "A"–top 25%; "B"–next 35%; "C"–next 30%; "D"—next 7%; "F"—bottom 3% (NYC Department of Education, 2011b). The results based on Scheme 3, which are largely consistent with those of Scheme 2, are available in Appendix C of the supplementary materials available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:13360004.

To distinguish the effects of the second and third factors—inverting the TCC, versus setting LOSS and HOSS arbitrarily—we replicated our earlier analyses while excluding students with arbitrarily set scores. Very few students were assigned the LOSS as the TC scale score: less than 0.2% from each combination of grade, subject, and year. Therefore, we focused on the effect of the HOSS on the two sets of scale scores, re-computing the effects of switching between the TC and Alt1-1 scales, using a restricted sample that excluded all students with perfect scores. The resulting effects reflect the contributions of other differences between the TC and Alt1-1 scales, net of the contribution of the difference in the HOSS.

Results

Severity of Ceiling Effects

The raw-score distributions of the results for the grades and years that we used in the study showed severe but varying ceiling effects. Skewness ranged from – 1.33 to – 0.53 (median = – 1.02) across subjects, grades, and years (column labeled "Raw" in Table 2). In general, ceiling effects were more severe in ELA (skewness from – 1.33 to – 1.05) than in mathematics (– 0.98 to – 0.53).

~~~~~ INSERT **TABLE 2** ABOUT HERE ~~~~~

These ceiling effects are comparable in severity to those observed for high-stakes tests in other states. For example, in their review of state-level scale score distributions from 17 states in 2010, Ho and Yu (2012) reported states with median skewness ranging from – 0.64 to – 0.87 (minimum = – 1.81) across grades 3-8 in reading and mathematics.

Differences Between the Two Scaling Approaches in Addressing Ceiling Effects

Both sets of scale scores lessen the ceiling effects by stretching out the upper tails of the raw-score distributions, but in most instances, the transformation imposed by the scaling
~~~~~

approach for the TC scale was considerably greater. This can be seen in two ways. First, the distributions of the Alt scale scores remained negatively skewed, with a median skewness of – 0.36 in ELA and – 0.28 in mathematics (panel I in Table 2). In contrast, in many instances, the skewness of the TC scale-score distribution was *positive*, with a median skewness of +1.94 in ELA and +0.13 in mathematics. Secondly, the distance between the median and the maximum TC scale scores was larger than that for the Alt scale scores (Table 3). For example, for grade-7 ELA in 2009, the distance between the median and the maximum TC scale scores was 4.19 SD while that for the Alt scale scores was 1.41 SD.

~~~~~ INSERT **TABLE 3** ABOUT HERE ~~~~~

For each combination of grade and year, the difference between the two scaling approaches in the upper-tail stretching is also consistently larger for ELA than that for mathematics (Table 3). In ELA, the differences between the two sets of scale scores in the distance from the median to the maximum scores on each set ranged from 2.45 SD to 3.13 SD. The corresponding differences in mathematics ranged from 0.99 SD to 1.55 SD. These results are consistent with the view that the two scaling approaches differ more in their stretching of the upper tail when the ceiling effects are more severe, as the ceiling effects were generally more severe in ELA.

The scaling approach for the TC scale scores also stretched the lower tails of the raw-score distributions more than that for the Alt scale scores. As a result, combining the effects of stretching at both ends of the distributions, the TC scale-score distributions have much larger kurtosis than the corresponding Alt scale-score distributions, the latter being—as expected—closer to a standard normal distribution. This applies to all combinations of subject, grade, and year. The median kurtosis for the TC scale scores was 10.68 in ELA and 5.81) in mathematics
~~~~~

(panel II in Table 2). In contrast, those for the Alt scale scores were smaller—medians of 3.07 in ELA and 2.89 in mathematics—and were close to that for a standard normal distribution (kurtosis = 3.00).

The above differences between the two scaling approaches do not appear to be driven simply by the fact that the TC scale scores are summed scores with a 1-1 mapping to the observed raw scores but the Alt scale scores are not. This is because, the skewness and kurtosis of the Alt1-1 scale scores—which are simply the summed-score version of the Alt scale scores—are comparable to those of the corresponding Alt scale scores rather than those of the TC scale scores (Table 2).

Estimated Impact on Schools' Performance Ratings

In this and the next two sections, we report our findings based on the model that included all control predictors (i.e., CA6). In a later section, we show that these findings are largely independent of the choice of model.

Impact on Schools' Ranks. Schools' ranks differed modestly between the two sets of scale scores. Across all combinations of grade, subject, and year, the Spearman's *rho* between school-performance estimates derived from the two sets of scale scores ranged from .89 to .98 (median = .97) (Table 4). Although these correlations are very high by usual conventions, they correspond to appreciable shifts in ranks. The average shift in ranks in either direction ranged from 132 to 50 rank positions (median = 69) on a listing of 1,243 schools.[8] That there were substantial shifts in ranks in some instances is also evident from the scatterplots of schools' ranks

[8] Other distributional statistics of the differences in school ranks (minimum, quartiles and maximum) are available in Appendix D of the supplementary materials available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:13360004.

based on the TC scale scores against their ranks based on the Alt scale scores, shown for the cases with the minimum and maximum Spearman's *rho* in Figure 2.

~~~~~ INSERT **TABLE 4** ABOUT HERE ~~~~~

For each grade and year, the scaling-approach effects tended to be larger for ELA than for mathematics (Table 4). For example, at grade 7 in 2009, the Spearman's *rho* for ELA was .93, corresponding to average shift in ranks in either direction of 104 rank positions, while that for mathematics was.97 (61 rank positions).

~~~~~ INSERT **FIGURE 2** ABOUT HERE ~~~~~

Impact on Classification of Schools in Broad Performance Bands. The scaling approach used affects schools' assigned performance bands, and as expected, the size of the impact depends on both the number and locations of the cut-scores. In Table 5, we display the observed percentages agreement between schools' assigned performance bands based on the two sets of scale scores, by grade, year, subject, and classification scheme, with chance agreement rates in parentheses and Cohen's *kappa* in italics.

~~~~~ INSERT **TABLE 5** ABOUT HERE ~~~~~

For performance bands defined by cut scores at ±1 posterior standard deviation (Scheme 1), the observed percentage agreement in classification was generally high, averaging 89% across grades, years, and subjects. However, chance agreement rates were also high, so Cohen's *kappa* ranged from moderate (.58) to high (.87), with a median of .65 for ELA and .85 for mathematics.

All other things being equal, the larger the number of cut scores, the higher the proportion of instances in which schools change classifications, and as expected, we found that the effect of switching from one scaling approach to another was larger with Scheme 2
~~~~~

(quintiles) than with Scheme 1. The average observed percentage agreement for Scheme 2 was 72%, with median Cohen's *kappa* of .56 for ELA and .74 for mathematics.

The Impact of the Severity of Ceiling Effects

The effect of switching from one scaling approach to another was substantially larger in cases where the ceiling effect on the raw-score distribution was more severe. For example, in the three cases where the raw-score distribution was the most negatively skewed—namely, grade-7 ELA in 2009 and 2010, and grade-8 ELA in 2010, with skewness ranging from – 1.33 to – 1.26—the average shift in ranks in either direction ranged from 103 to 132 rank positions. These are considerably larger than the corresponding average shift in ranks in either direction, ranging from 55 to 62 rank positions, observed for the three cases with the least negatively skewed raw-score distributions—namely, grade-7 mathematics in 2009 and 2010, and grade-8 mathematics in 2010, with skewness ranging from – 0.83 to – 0.53.

Relative Importance of Different Scaling Decisions to the Scaling-Approach Effect

As expected, the use of summed scores—that is, a 1-to-1 mapping between scale scores and raw scores—did not contribute substantially to the scaling-approach effects on schools' ratings. Switching between the TC scale scores and the Alt1-1 scale scores—the 1-to-1 version of the Alt scale scores—had virtually the same effect as switching between the TC scale scores and the Alt scale scores. This applies to both school ranks and performance bands, and to different combinations of grade, subject, and year (Table 6 with Tables 4 and 5). The only exception was for grade-7 mathematics in 2010. Although the Spearman's *rho* for the TC versus Alt1-1 scale scores (.99) was very close to that for the TC versus Alt scale scores (.98), for reasons that are unclear, the difference in average shift in ranks between the two sets of results (20 rank positions) was slightly larger than that observed for the other combinations of grade,

subject, and year (between 0 and 11 rank positions). Nonetheless, even in this case, the difference can be considered small on a rank list of 1,243 schools.

~~~~~ INSERT **TABLE 6** ABOUT HERE ~~~~~

As noted, the effect of switching from one scaling approach to the other was larger when the ceiling effect was more severe, and in those cases, the difference in the location of the HOSS between the two scaling approaches was a key driver of the scale-approach effects. Specifically, in the three cases where the raw-score distribution was the most negatively skewed among those included in the study—namely, grade-7 ELA in 2009 and 2010, and grade-8 ELA in 2010—the exclusion of the perfect scorers led to considerably more consistent school-performance estimates between the TC scale scores and the Alt1-1 scale scores than those obtained in the full analytic sample (compare panels I and II in Table 6). For example, for grade-7 ELA in 2009, the average shift in ranks in either direction was 98 and 48 rank positions in the full analytic sample and the restricted sample respectively.

In contrast, in cases where the raw-score distribution was less negatively skewed (e.g., grade-7 mathematics in 2009 and 2010, and grade-8 mathematics in 2010), using the restricted sample did not lead to substantial reduction in the effects. This could be due to (1) the already very small scaling-approach effects in the full analytic sample in these cases, which might create a floor effect on the scaling-approach effects; or (2) the relatively small number of perfect scorers in these cases; or both.

Sensitivity of Results to Choice of Model

The estimated scaling-approach effects on both schools' ranks and performance bands were largely independent of the type of covariates included in the model. There is limited variation in Spearman's *rho* among the six models (Table 4): the range across models was only
~~~~~

from .01 to .06 (median = .01), depending on grade, subject, and year. The corresponding range of the average shift in ranks in either direction was between 6 and 43 rank positions (median = 10), with larger variation observed for ELA (between 11 and 43 rank positions) than for mathematics (between 6 and 9 rank positions). Similarly, there is also minimal variation among the six models in percentage agreement in schools' assignment to performance bands (Table 5): the range across models was between 1% and 12% (median = 4%), depending on grade, subject, year, and classification schemes.

Similarly, the other results—(1) the scaling-approach effects were larger where the ceiling effects on the raw-score distributions were more severe; (2) the use of summed scores *per se* contributed little to the scaling-approach effects; and (3) the difference in the location of the HOSS between the two sets of scale scores was the key driver of the scaling-approach effects in cases with severe ceiling effects on the raw-score distributions—all hold for each of the other models. Details are available upon request.

Discussion

In this study, we found that the choice between two scaling methods—a commonly used method for creating summed scores, and an alternative Bayesian approach—affected schools' performance ratings modestly. Scaling affected both schools' ranks and their assignment to broad performance bands. These effects were larger when the underlying raw score distribution had more severe ceiling effects, and in such cases, the inconsistency in schools' ratings was primarily driven by the difference in the location of the HOSS on the two sets of scale scores.

Although the effects are modest in size, they are large enough to be important when there are high stakes attached to school ratings. As both approaches we used are conventional and reasonable alternatives, and because the choice between them is substantively unrelated to the

inference, these inconsistencies pose a threat to the validity of school ratings. Ceiling effects, which drive some of the inconsistency we found, are very common in current high-stakes testing programs, and methods similar to those used to derive the TC scale scores, are used in several states. Moreover, policymakers, educators, or the public typically do not understand the methodological choices involved, and the inconsistencies inherent in the choices among scaling options are generally not revealed to them.

While we contrasted only two approaches, other research suggests that the problem may be more general. For example, Kim and Nicewander (1993) demonstrated that a maximum-likelihood estimator without 1-to-1 mapping to raw scores, a weighted-likelihood estimator, a Bayesian-modal estimator, and an EAP estimator perform comparably in the middle ability ranges but differ at the ends of the ability spectrum. Insofar as high- and low-ability students are distributed unevenly among schools, choosing a different score estimator among this list could affect schools' ratings. These differences would be particularly pertinent in the context of high-stakes testing in view of the impact of ceiling effects.

Policymakers and test designers can take steps to reduce the uncertainty that we found. Score distributions should be monitored for ceiling effects, and when severe ceiling effects are detected, more difficult items should be added to subsequent tests. Our results show that while the adoption of a 1-to-1 mapping *per se* does not matter substantially, the method of obtaining such a mapping does, and is in fact the primary driver of the impact that we found. In particular, if testing contractors use methods that set the LOSS or HOSS *a priori*, the effects of these on the distribution of the scale scores should be carefully monitored, and HOSS and LOSS values distant from the rest of the distribution should be avoided if possible. This is particularly important in the presence of ceiling effects on the raw-score distributions.

However, even if steps are taken to lessen the inconsistencies found here, test-based school ratings may remain sensitive to reasonable alternative choices about the construction and scaling of the test. The risks of substantial inconsistencies are greatest when some schools have disproportionately large proportions of students with extreme scores. The best way to address these threats to validity may lie outside of testing—for example, using additional data, along with scores, to evaluate schools' performance.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351-383.

Briggs, D. C., & Betebenner, D. W. (2009, April). *Is growth in student achievement scale dependent?* Paper Presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy, 4*(4), 384-414.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011, March). *Teacher effectiveness on high- and low-stakes tests.* Paper presented at the Society for Research on Educational Effectiveness Spring 2011 Conference, Washington, D.C.

Crone, L. J., Lang, M. H., & Teddlie, C. (1995). Achievement measures of school effectiveness: Comparison of model stability across years. *Applied Measurement in Education, 8*(4), 353-365.

CTB/McGraw-Hill. (2006). *New York State testing program 2006: English language arts, grades 3-8 technical report*. New York: New York State Education Department.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement* (Working Paper 06-1). Policy Analysis for California Education, PACE.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment Research & Evaluation, 17*(1), May 25, 2012.

Ho, A., & Yu, C. (2012, under review). Descriptive statistics for modern test score distributions: Beyond skewness and kurtosis, *Educational and Psychological Measurement*.

Jacob, B. A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments* (NBER Working Paper No. 12817). National Bureau of Economic Research.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*(4), 587-99.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*(1), 54-81.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*(3), 245-56.

McKinley, R. L., & Reckase, M. D. (1980). *A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness*

of fit as a criterion (research report) No. 80-2). Columbia, MO: University of Missouri, Educational Psychology Department, Tailored Testing Research Laboratory.

National Research Council. (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academes Press.

New York City Department of Education. (2011a). *Educator guide: The New York City Progress Report Elementary/Middle/K-8 (2009-10).* Retrieved March 9, 2012, from http://schools.nyc.gov/NR/rdonlyres/4015AD0E-85EE-4FDE-B244-129284A7C36C/0/EducatorGuide_EMS_2011_03_10.pdf

New York City Department of Education. (2011b). *Educator guide: The New York City Progress Report Elementary/Middle/K-8 (2010-11)*. Retrieved March 9, 2012, from http://schools.nyc.gov/NR/rdonlyres/A82481C5-A351-47BA-BF8C-9F353E9CFB22/0/EducatorGuide_EMS_2011_10_03.pdf

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi:10.3102/0002831210362589

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis method*s (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item Response Theory for Items Score in More than Two Categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, N.J.: Lawrence Erlbaum Associates.

Thissen, D., & Orlando, M. (2001). Item Response Theory for Items Score in Two Categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, N.J.: Lawrence Erlbaum Associates.

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 100-109). Thousand Oaks, CA: Corwin.

Appendix A. Name, definition and coding of the principal variables

SN	Variable	Description
I. Outcome Variables (Student level)		
1.	*TC_MATH(Y)*	Mathematics scale score provided by testing company
2.	*TC_ELA(Y)*	English Language Arts scale score provided by testing company
3.	*Alt_MATH(Y)*	Mathematics scale score created by The Harvard Education Accountability Project
4.	*Alt_ELA(Y)*	English Language Arts scale score created by The Harvard Education Accountability Project
II. Previous Year's Achievement Variables (Student level)		
5.	*TC_MATH(Y-1)*	Mathematics scale score on NYS mathematics test in immediate previous grade from grade in target year provided by testing company
6.	*TC_ELA(Y-1)*	English Language Arts scale score on NYS mathematics test in immediate previous grade from grade in target year provided by testing company
7.	*Alt_MATH(Y-1)*	Mathematics scale score on NYS mathematics test in immediate previous grade from grade in target year created by The Harvard Education Accountability Project
8.	*Alt_ELA(Y-1)*	English Language Arts scale score on NYS mathematics test in immediate previous grade from grade in target year created by The Harvard Education Accountability Project
III. Milestone Grade's Achievement Variables (Student level)		
9.	*TC_MATH(PA)*	Mathematics scale score on NYS mathematics test in grade 5 provided by testing company
10.	*TC_ELA(PA)*	English Language Arts scale score on NYS mathematics test in grade 5 provided by testing company
11.	*Alt_MATH(PA)*	Mathematics scale score on NYS mathematics test in grade 5 created by The Harvard Education Accountability Project
12.	*Alt_ELA(PA)*	English Language Arts scale score on NYS mathematics test in grade 5 created by The Harvard Education Accountability Project
IV. Background Variables (Student level)		
13.	*SCHOOLID*	School student belongs to at point of taking state test
14.	*GRADE*	Grade level of state test taken in target year
15.	*FEMALE*	Binary variable coding for student's gender
16.	Race/Ethnicity	A set of binary variables coding for student's race/ethnicity
17.	*LOW_INC*	Binary variable coding for student's family income status
18.	*IMMIGRANT*	Binary variable coding for student's immigrant status
19.	*LEP*	Binary variable coding for Limited-English-Proficiency status
20.	*DISABLED*	Binary variable coding for disability status
21.	*ACCOMMOD*	Binary variable coding for whether student received accommodations for the state test

Table 1

Classification and model specification of school-performance measures, by type of covariates

Type of Covariates	**Model**
1. None	CA1
2. Student-level background only	CA2
3. Student- and school-level background	CA3
4. Student-level prior achievement in milestone grade only	CA4
5. Student-level prior achievement in milestone grade and background	CA5
6. Student- and school-level prior achievement in milestone grade and background	CA6

Table 2

Skewness and kurtosis of raw-score distributions and three scale-score distributions in the analytic sample, by subject, grade, and year

	ELA				Math			
Year	Raw	TC Scale	Alt Scale	Alt1-1 Scale	Raw	TC Scale	Alt Scale	Alt1-1 Scale
I. Skewness								
A. Grade 7								
2009	– 1.33	1.87	– 0.44	– 0.45	– 0.74	0.71	– 0.28	– 0.29
2010	– 1.28	2.23	– 0.36	– 0.38	– 0.53	– 0.19	– 0.21	– 0.19
B. Grade 8								
2009	– 1.05	1.16	– 0.30	– 0.33	– 0.98	– 0.03	– 0.38	– 0.39
2010	– 1.26	2.01	– 0.35	– 0.42	– 0.83	0.28	– 0.27	– 0.28
II. Kurtosis								
A. Grade 7								
2009	4.86	11.62	3.10	3.07	2.84	6.56	2.86	2.83
2010	4.67	10.56	2.87	2.87	2.69	6.81	2.92	2.84
B. Grade 8								
2009	4.14	8.49	3.04	3.02	3.22	5.05	3.01	2.97
2010	4.95	10.80	3.11	3.25	2.88	5.03	2.84	2.82

Table 3

Distance from the median standardized score to the maximum standardized score for the two set of scale scores in the analytic sample, by grade, year, and subject

	ELA			Mathematics		
Year	TC scale	Alt scale	Difference (TC – Alt)	TC scale	Alt scale	Difference (TC – Alt)
A. Grade 7						
2009	4.19	1.41	2.78	3.18	1.63	1.55
2010	4.33	1.20	3.13	3.21	1.97	1.24
B. Grade 8						
2009	4.33	1.88	2.45	2.94	1.95	0.99
2010	4.46	1.45	3.01	2.94	1.80	1.14

Table 4

Spearman's rho between school-performance estimates derived from the TC and Alt scale scores (mean absolute difference in ranks in parentheses), by grade, year, and subject

	Model CA6		Median for Models CA1 to CA5	
Grade/Year	**ELA**	**Mathematics**	**ELA**	**Mathematics**
A. Grade 7				
2009	.93 (104)	.97 (61)	.93 (98)	.98 (57)
2010	.89 (132)	.98 (56)	.90 (119)	.98 (52)
B. Grade 8				
2009	.96 (77)	.98 (50)	.96 (72)	.98 (56)
2010	.92 (104)	.98 (55)	.94 (93)	.98 (56)

Table 5

Observed percentage agreement between school-performance estimates derived from the TC and Alt scale scores (chance agreement rates in parentheses, Cohen's kappa in italics), by grade, year, subject, and classification scheme

Grade /Year	**Model CA6**				**Median for Models CA1 to CA5**	
	ELA		**Mathematics**		**ELA**	**Mathematics**
	Scheme 1	**Scheme 2**	**Scheme 1**	**Scheme 2**	**Scheme 1**	**Scheme 1**
A. Grade 7						
2009	84 (56)	64 (20)	91 (51)	78 (20)	86 (56)	90 (51)
	.63	*.55*	*.83*	*.73*	*.69*	*.80*
2010	81 (55)	55 (20)	93 (53)	80 (20)	85 (55)	93 (52)
	.58	*.43*	*.86*	*.75*	*.64*	*.86*
B. Grade 8						
2009	89 (54)	71 (20)	93 (52)	81 (20)	90 (54)	92 (52)
	.75	*.64*	*.87*	*.77*	*.78*	*.84*
2010	85 (55)	65 (20)	92 (53)	79 (20)	87 (55)	92 (53)
	.67	*.56*	*.84*	*.72*	*.72*	*.83*

Table 6

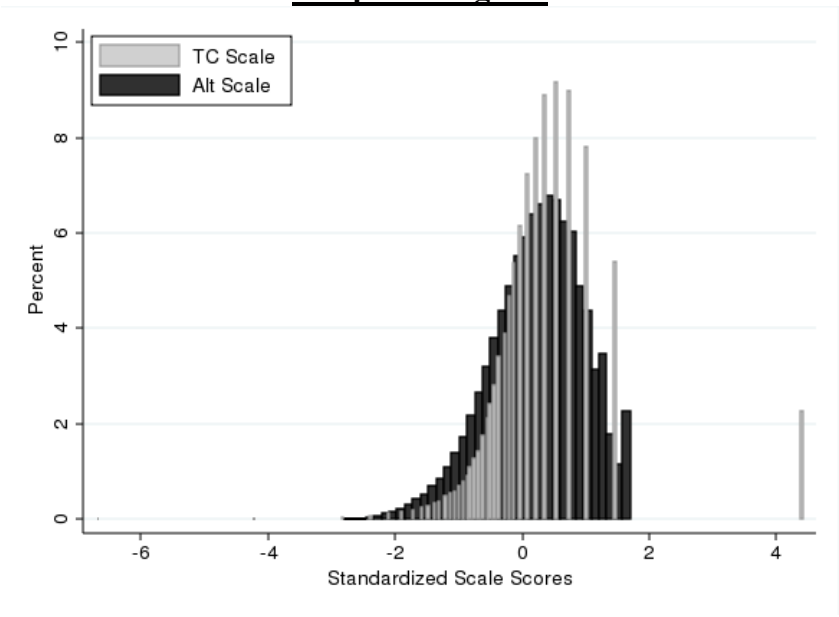
Estimated scaling-approach effects due to a switch in the TC scale scores and Alt1-1 scale scores, using model CA6, by grade, year, and subject

	Spearman's Rank Correlation (Mean Absolute Difference in Ranks)		Percentage Agreement for Scheme 1	
Grade / Year	**ELA**	**Mathematics**	**ELA**	**Mathematics**
(I) Full Analytic Sample				
A. Grade 7				
2009	.94 (98)	.98 (54)	85	92
2010	.89 (132)	.99 (36)	81	96
B. Grade 8				
2009	.96 (72)	.99 (41)	89	95
2010	.93 (101)	.99 (44)	85	94
(II) Restricted Sample Excluding Perfect Scorers				
A. Grade 7				
2009	.98 (48)	.99 (37)	93	95
2010	.98 (47)	.99 (32)	94	97
B. Grade 8				
2009	.98 (52)	.99 (38)	93	95
2010	.98 (53)	.99 (36)	92	95

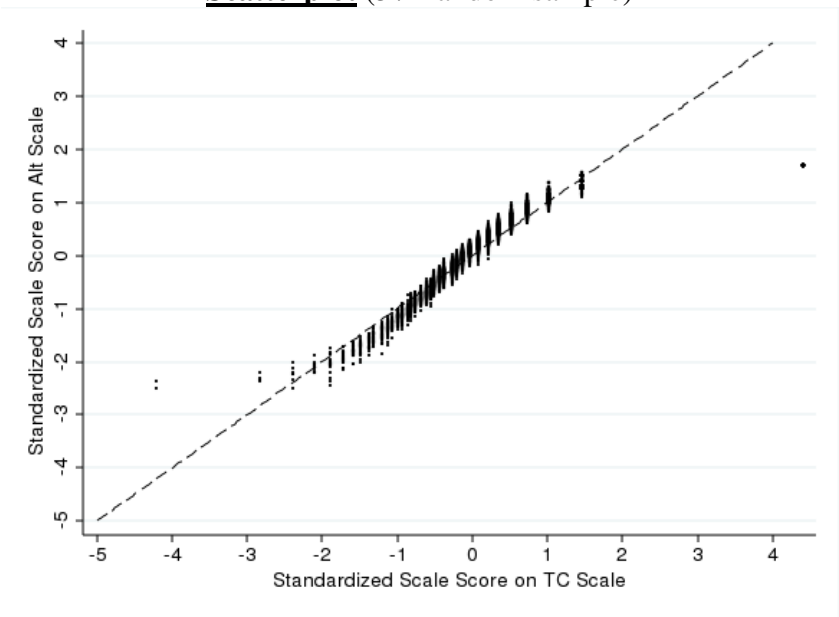
Figure 1

Comparisons of sample distributions of standardized TC and Alt scale scores for 2009 grade-7 ELA

Sample Histogram



Scatterplot (5% random sample)



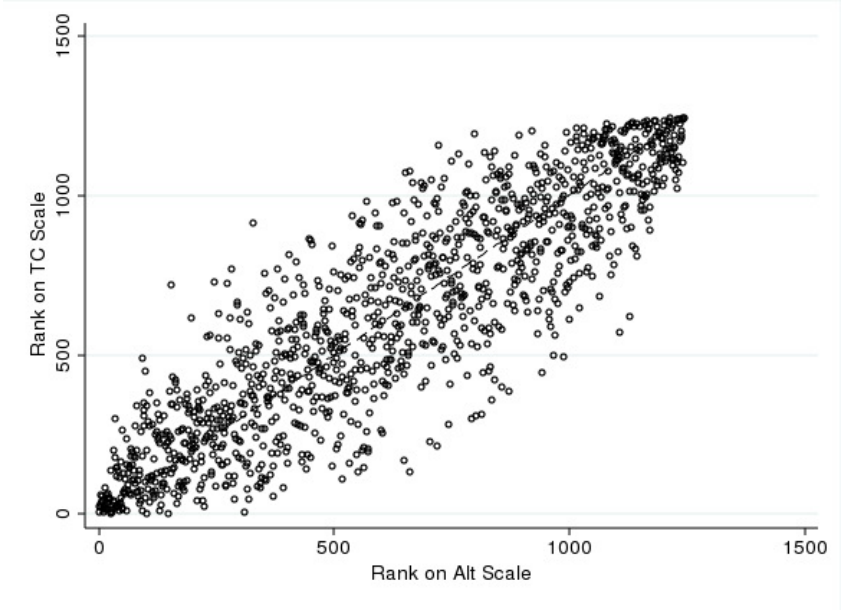
Note. The dashed line on the scatter-plot on the bottom panel is the identity line, i.e., standardized Alt scale score = standardized TC scale score.

Figure 2

Scatterplots of schools' ranks on the TC and Alt scale scores, for the minimum and maximum Spearman's rho (model CA6)

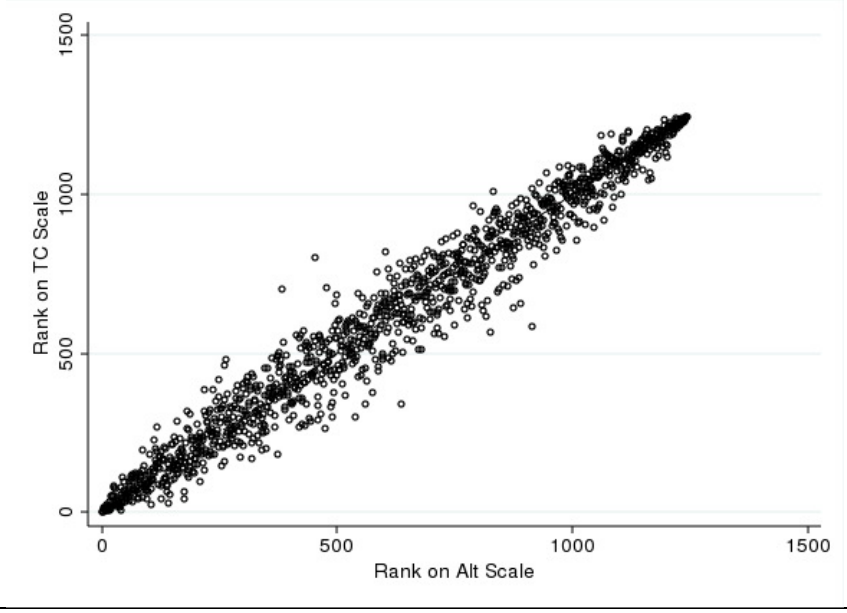
Minimum

Grade 7 2010 ELA (correlation = .89; MAD = 132)



Maximum

Grade 8 2009 Mathematics (correlation = .98; MAD = 50)



Note. The dashed line on each scatter-plot is the identity line, i.e., rank on TC scale = rank on Alt scale.

Supplementary Materials

Appendix A: Scoring Methods Adopted by 36 States

State	Year of Technical Report	1-to-1 Mapping Between Raw Scores and Scale scores		Not 1-to-1 Mapping	Source
		Rasch	Non-Rasch		
Alaska	2011	✓			http://www.eed.state.ak.us/tls/assessment/techreports.html
Arizona	2011	✓			http://www.azed.gov/standards-development-assessment/files/2011/12/aims_tech_report_2011_final.pdf
California	2011	✓			http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt2010.pdf
Colorado	2011			✓	http://www.cde.state.co.us/cdeassess/publications.html
Delaware	2008	✓			http://www.doe.k12.de.us/aab/files/tech_report_2008.pdf#DSTP Technical Report 2008
District of Columbia	2011		✓		http://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/DC_CAS_2011_technical_report_v6%20for%20OSSE%2006-30-11.pdf
Florida	2006			✓	http://fcat.fldoe.org/pdf/fc06tech.pdf
Idaho	2011	✓			http://www.sde.idaho.gov/site/assessment/ISAT/docs/technicalReports/ISAT%20Spring%202011%20Technical%20Report_Final.pdf
Illinois	2011		✓		http://www.isbe.net/assessment/htmls/isat_general_info.htm#tech
Indiana	2011			✓	http://www.doe.in.gov/sites/default/files/assessment/gti.pdf
Maine*	2011		✓		http://www.maine.gov/education/mea/1011materials/techincal_report.pdf
Maryland	2007	✓			http://www.marylandpublicschools.org/MSDE/divisions/planningresultstest/2007+MSA+Reading+Technical+Report
Massachusetts	2010		✓		http://www.doe.mass.edu/mcas/tech/?section=techreports
Michigan	2011	✓			http://www.michigan.gov/documents/mde/MEAP_FALL_2011_Guide_to_Reports_377049_7.pdf
Missouri	2010			✓	http://dese.mo.gov/divimprove/assess/tech/
Montana	2011	✓			http://www.opi.mt.gov/PDF/Assessment/CRT/10-11Montana-CRT-Tech-Report.pdf
Nebraska	2011	✓			http://www.education.ne.gov/assessment/pdfs/Final_NeSA_2011_Technical_Report.pdf

Continue on next page...

...continued from previous page

State	Year of Technical Report	1-to-1 Mapping Between Raw Scores and Scale scores		Not 1-to-1 Mapping	Source
		Rasch	Non-Rasch		
New Hampshire*	2011		✓		http://www.education.nh.gov/instruction/assessment/necap/documents/techrpt_july2011.pdf
New Jersey	2010	✓			http://www.state.nj.us/education/assessment/es/njask_tech_report09.pdf
New Mexico	2011		✓		http://ped.state.nm.us/AssessmentAccountability/AssessmentEvaluation/dl11/2010_11_NM_SBA_Tech_Report.pdf
New York	2011		✓		http://www.p12.nysed.gov/apda/reports/
North Carolina	2009		✓		http://www.ncpublicschools.org/accountability/testing/technicalnotes
North Dakota	2011		✓		http://www.dpi.state.nd.us/testing/assess/10final.pdf
Ohio	2011	✓			http://www.education.ohio.gov/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1143&ContentID=9479&Content=117721
Oklahoma	2009		✓		http://www.ok.gov/sde/sites/ok.gov.sde/files/2009Gr3_8.pdf
Oregon	2008	✓			http://www.ode.state.or.us/search/page/?=1305
Pennsylvania	2011	✓			http://www.portal.state.pa.us/portal/server.pt/community/technical_analysis/7447
Rhode Island*	2011		✓		http://www.ride.ri.gov/Assessment/DOCS/NECAP/Tech_Manual/Archive/2010-11_NECAP_Math-Reading-Writing_Technical_Report_with_Appendices.pdf
South Carolina	2010	✓			http://ed.sc.gov/agency/ac/Assessment/AssessmentTechnicalReports.cfm
Texas	2010	✓			http://www.tea.state.tx.us/student.assessment/techdigest/
Vermont*	2011		✓		http://www.ride.ri.gov/Assessment/DOCS/NECAP/Tech_Manual/Archive/2010-11_NECAP_Math-Reading-Writing_Technical_Report_with_Appendices.pdf
Virginia	2009	✓			http://www.doe.virginia.gov/testing/test_administration/technical_reports/sol_technical_report_2008-09_administration_cycle.pdf
Washington	2010	✓			http://www.k12.wa.us/assessment/pubdocs/WCAP_2010SpringAdmin_TechReport.pdf

Continue on next page...

...continued from previous page

State	Year of Technical Report	1-to-1 Mapping Between Raw Scores and Scale scores		Not 1-to-1 Mapping	Source
		Rasch	Non-Rasch		
West Virginia	2008			✓	http://wvde.state.wv.us/oaa/pdf/WESTEST%202008%20Supplemental%20Report%20_3_.pdf
Wisconsin	2010			✓	http://dpi.wi.gov/oea/pdf/td-2010-techman.pdf
Wyoming	2011	✓			http://www.edu.wyoming.gov/Libraries/Assessments/PAWS_2010-2011_Technical_Manual_gwg_draft.sflb.ashx
	Total	18	12	6	

*Information for these four states is for the "New England Common Assessment Program" that the states collaborated to develop and use for students in grades 3 to 8 and 11 to meet the *No Child Left Behind* requirements.

Appendix B. Descriptive Statistics of the Total and Analytic Sample

Table 1

Student-level descriptive statistics of total and analytic samples

Variables	**Total Sample**		**Analytic Sample**	
No. of schools (*N*)	1,451		1,243	
No. of students (*n*)	765,843		661,504	
Grade				
% Grade 7	49		51	
% Grade 8	51		49	
Year				
% 2009	50		50	
% 2010	50		50	
	Mean	**SD**	**Mean**	**SD**
I. Outcome Variables				
TC_MATH(Y)	0.09	0.91	0.14	0.88
TC_ELA(Y)	0.09	1.01	0.14	1.00
Alt_MATH(Y)	0.13	0.90	0.18	0.87
Alt_ELA(Y)	0.11	0.84	0.17	0.80
II. Previous Year's Achievement Variables				
TC_MATH(Y-1)	0.19	0.90	0.25	0.88
TC_ELA(Y-1)	0.11	0.90	0.16	0.88
Alt_MATH(Y-1)	0.23	0.90	0.29	0.87
Alt_ELA(Y-1)	0.10	0.80	0.16	0.76

Continue on next page...

Table 1 (continued)

Student-level descriptive statistics of total and analytic samples

Variables	**Total Sample**		**Analytic Sample**	
	Mean	**SD**	**Mean**	**SD**
III. Milestone Grade's Achievement Variables				
TC_MATH(PA)	0.24	0.93	0.27	0.92
TC_ELA(PA)	0.10	0.88	0.12	0.87
Alt_MATH(PA)	0.25	0.92	0.28	0.91
Alt_ELA(PA)	0.03	0.86	0.04	0.85
IV. Background Variables				
FEMALE	0.49	-	0.50	-
Race/Ethnicity				
WHITE	0.54	-	0.56	-
BLACK	0.18	-	0.17	-
HISPANIC	0.20	-	0.18	-
ASIAN	0.07	-	0.07	-
INDIAN	0.00	-	0.00	-
PACISLDER	0.00	-	0.00	-
MULTIRACIAL	0.00	-	0.00	-
LOW_INC	0.48	-	0.46	-
IMMIGRANT	0.02	-	0.01	-
LEP	0.04	-	0.03	-
DISABLED	0.14	-	0.13	-
ACCOMMOD	0.18	-	0.16	-

Table 2

School-level descriptive statistics of total and analytic samples

Variables	**Total Sample**		**Analytic Sample**	
	Mean	**SD**	**Mean**	**SD**
I. Outcome Variables				
SCHM_TC_MATH(Y)	0.08	0.42	0.14	0.40
SCHM_TC_ELA(Y)	0.08	0.39	0.14	0.37
SCHM_Alt_MATH(Y)	0.12	0.44	0.18	0.41
SCHM_Alt_ELA(Y)	0.11	0.37	0.17	0.34
II. Previous Year's Achievement Variables				
SCHM_TC_MATH(Y-1)	0.18	0.40	0.25	0.37
SCHM_TC_ELA(Y-1)	0.10	0.35	0.16	0.33
SCHM_Alt_MATH(Y-1)	0.22	0.42	0.29	0.39
SCHM_Alt_ELA(Y-1)	0.10	0.34	0.16	0.31
III. Milestone Grade's Achievement Variables				
SCHM_TC_MATH(PA)	0.23	0.40	0.27	0.38
SCHM_TC_ELA(PA)	0.09	0.35	0.12	0.34
SCHM_Alt_MATH(PA)	0.23	0.40	0.28	0.39
SCHM_Alt_ELA(PA)	0.01	0.37	0.04	0.36

Continue on next page...

Table 2 (continued)

School-level descriptive statistics of total and analytic samples

Variables	**Total Sample**		**Analytic Sample**	
	Mean	**SD**	**Mean**	**SD**
IV. Background Variables				
SCHM_FEMALE	0.49	0.04	0.50	0.04
Race/Ethnicity				
SCHM_WHITE	0.54	0.38	0.56	0.38
SCHM_BLACK	0.18	0.24	0.17	0.24
SCHM_HISPANIC	0.20	0.24	0.18	0.23
SCHM_ASIAN	0.07	0.12	0.07	0.12
SCHM_INDIAN	0.00	0.02	0.00	0.02
SCHM_PACISLDER	0.00	0.00	0.00	0.00
SCHM_MULTIRACIAL	0.00	0.01	0.00	0.00
SCHM_LOW_INC	0.48	0.34	0.46	0.34
SCHM_IMMIGRANT	0.02	0.04	0.01	0.03
SCHM_LEP	0.04	0.07	0.03	0.05
SCHM_DISABLED	0.14	0.05	0.13	0.05
SCHM_ACCOMMOD	0.18	0.08	0.16	0.07

Appendix C. Results for Classification Scheme 3 (Unequal Proportion, Asymmetric)

Table 1

Observed percentage agreement between school-performance estimates derived from the TC and Alt scale scores, model CA6, for classification Scheme 3 (chance agreement rates in parentheses, Cohen's kappa in italics), by grade, year, and subject

Grade / Year	ELA	Mathematics
A. Grade 7		
2009	72 (28)	82 (28)
	.62	*.74*
2010	62 (28)	86 (28)
	.48	*.80*
B. Grade 8		
2009	79 (28)	86 (28)
	.70	*.80*
2010	71 (28)	86 (28)
	.61	*.79*

Appendix D. Differences in School Ranks based on the TC and Alt Scale Scores

Table 1

Distribution of difference in rank between school-performance estimates derived from the TC and Alt scale scores, using model CA6, by grade, year, and subject

Grade/Year	Mean	Min	25th Percentile	50th Percentile	75th Percentile	Max
I. ELA						
A. Grade 7						
2009	0	- 447	- 87	- 1	73	585
2010	0	- 586	- 106	- 6	107	535
B. Grade 8						
2009	0	- 366	- 58	- 4	56	550
2010	0	- 558	- 80	- 4	73	576
II. Mathematics						
A. Grade 7						
2009	0	- 317	- 48	- 1	46	409
2010	0	- 384	- 41	1	41	275
B. Grade 8						
2009	0	- 348	- 39	- 1	34	330
2010	0	- 294	- 44	0	43	295