Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 16-09

Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013

AUTHORS

Sean F. Reardon
Stanford University

Demetra Kalogrides
Stanford University

Andrew D. Ho
Harvard University

ABSTRACT

In the U.S., there is no recent database of district-level test scores that is comparable across states. We construct and evaluate such a database for years 2009-2013 to support large-scale educational research. First, we derive transformations that link each state test score scale to the scale of the National Assessment of Educational Progress (NAEP). Next, we apply these transformations to a unique nationwide database of district-level means and standard deviations, obtaining estimates of each districts' test score distribution expressed on the NAEP measurement scale. We then conduct a series of validation analyses designed to assess the validity of key assumptions underlying the methods and to assess the extent to which the districts' transformed distributions match the districts' actual NAEP score distributions (for a small subset of districts where the NAEP assessments are administered). We also examine the correlations of our estimates with district test score distributions on a second "audit test"—the NWEA MAP test, which is administered to populations of students in several thousand school districts nationwide. Our linking method yields estimated district means with a root mean square deviation from actual NAEP scores of roughly 1/10th of a standard deviation unit in any single year or grade. The correlations of our estimates with average district means over years and grades are .97-.98 for NAEP and 0.93 for the NWEA test. We conclude that the linking method is accurate enough to be used in large-scale educational research about national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions or rankings among districts in different states invalid.

VERSION

April 2016

Suggested citation: Reardon, S.F., Kalogrides, D., & Ho, A. (2016). Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013 (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09

Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013

Sean F. Reardon
Demetra Kalogrides
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

April, 2016

The research described here was supported by grants from the Institute of Education Sciences (R305D110018), the Spencer Foundation, and the William T. Grant Foundation. Some of the data used in this paper were provided by the National Center for Education Statistics (NCES). The paper would not have been possible without the assistance of Ross Santy, Michael Hawes, and Marilyn Seastrom, who facilitated access to the EdFacts data. Additionally, we are grateful to Yeow Meng Thum at NWEA, who provided the NWEA data used in some analyses. This paper benefitted substantially from ongoing collaboration with Erin Fahle, Ken Shores, and Ben Shear. The opinions expressed here are our own and do not represent views of NCES, NWEA, the Institute of Education Sciences, the Spencer Foundation, the William T. Grant Foundation, or the U.S. Department of Education. Direct correspondence and comments to Sean F. Reardon, sean.reardon@stanford.edu, 520 CERAS Building #526, Stanford University, Stanford, CA 94305.

Linking U.S. School District Test Score Distributions to a Common Scale, 2009-2013

Abstract

In the U.S., there is no recent database of district-level test scores that is comparable across states. We construct and evaluate such a database for years 2009-2013 to support large-scale educational research. First, we derive transformations that link each state test score scale to the scale of the National Assessment of Educational Progress (NAEP). Next, we apply these transformations to a unique nationwide database of district-level means and standard deviations, obtaining estimates of each districts' test score distribution expressed on the NAEP measurement scale. We then conduct a series of validation analyses designed to assess the validity of key assumptions underlying the methods and to assess the extent to which the districts' transformed distributions match the districts' actual NAEP score distributions (for a small subset of districts where the NAEP assessments are administered). We also examine the correlations of our estimates with district test score distributions on a second "audit test"—the NWEA MAP test, which is administered to populations of students in several thousand school districts nationwide. Our linking method yields estimated district means with a root mean square deviation from actual NAEP scores of roughly $1/10^{th}$ of a standard deviation unit in any single year or grade. The correlations of our estimates with average district means over years and grades are .97-.98 for NAEP and 0.93 for the NWEA test. We conclude that the linking method is accurate enough to be used in large-scale educational research about national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions or rankings among districts in different states invalid.

Introduction

U.S. school districts differ dramatically in their socioeconomic and demographic characteristics (Reardon, Yun, & Eitle, 1999; Stroub & Richards, 2013), and districts have considerable influence over instructional and organizational practices that may affect academic achievement (Whitehurst, Chingos, & Gallaher, 2013). Nonetheless, we have relatively little rigorous large-scale research describing national patterns of variation in achievement across districts, let alone an understanding of the factors that cause this variation. Such analyses generally require district-level test score distributions that are comparable across states. No such nation-wide, district-level achievement dataset currently exists. This paper proposes and evaluates a linking method to construct such a dataset for research purposes.

Table 1 shows the scope of selected assessment systems to illustrate the gap that we attempt to fill. At the coarsest level, the National Assessment of Educational Progress (NAEP) provides comparable state-level scores in odd years, in reading and mathematics, in grades 4 and 8. NAEP also provides district-level scores, but only for around 20 large urban districts under the Trial Urban District Assessment (TUDA) initiative. Within individual states, we can compare district achievement using state math and English Language Arts (ELA) tests federally mandated by the No Child Left Behind (NCLB) act, administered annually in grades 3-8.

Table 1 here

Comparing academic achievement across state lines requires either that districts administer a common test, or that the scores on the tests can be roughly linked. However, other than in four New England states that shared a common assessment for a time, state accountability tests differ across states. Each state develops and administers its own tests; these tests may assess somewhat different content domains; scores are reported on different, state-determined scales; and proficiency thresholds are set at different levels of achievement. Moreover, the content, scoring, and definition of proficiency may vary within any given state over time and across grades. As a result, comparing scores on state

accountability tests across states (or in many cases within states across grades and years) has not been possible. The development of common assessments developed by multistate assessment consortia (such as PARCC and SBAC) will certainly increase comparability across states, but only to the extent that states adopt these assessments. Given the incomplete, divided, and declining state participation in these consortia, the availability of comparable district-level test score data among all districts remains out of reach.

In some cases, districts also administer voluntarily-chosen assessments, often for formative assessment purposes. When two districts adopt the same such assessments, we can compare test scores on these assessments among districts. One of the most widely used assessments, the Measures of Academic Progress (MAP) test from Northwest Evaluation Association (NWEA), is voluntarily administered in several thousand school districts, over 20% of all districts in the country. Although this allows for comparison between NAEP and MAP scores in these districts, the districts using MAP are not a representative sample of districts.

In this paper, we propose and assess a method of rendering district-level average state accountability test scores comparable across states, years, and grades. We rely on a combination of a) representative state-level test score distributions from NAEP and population state test score data from every school district in the U.S.; b) the estimation of scale transformations that link state test scores to the NAEP scale; c) a set of validation checks to assess the accuracy of the resulting linked estimates; and d) three approaches to standardizing the resulting scale. None of the components of our method is novel on its own, but together they yield a national district-level dataset with considerable promise for enabling research.

We use data sources from the EDFacts Initiative (U.S. Department of Education, 2015), the NAEP Data Explorer, and NWEA. We estimate the necessary transformations using a) an application of heteroscedastic ordered probit (HETOP) models to transform district proficiency counts to standardized

district means and variances (Reardon, Shear, Castellano, & Ho, 2016); and b) linear test score linking methods (reviewed by Kolen and Brennan, 2014). Our validation checks rely on assessing the alignment of the linked district means to their respective NAEP TUDA and NWEA MAP distributions. We also standardize the linked scores to represent a) standardized scores relative to the year-, grade-, and subject-specific distribution of scores within a given cohort of students; b) standardized scores relative to the grade-and subject-specific distribution of scores within a given cohort of students; and c) scores expressed in units of average subject-specific grade-level differences of the national student population.

We are not the first to use methods of this sort to render scores on different tests comparable. Hanushek and Woessman (2012) used similar methods for country-level international comparisons. At the district level, Greene and McGee (2011) mapped 2004, 2005, 2007, and 2009 scores onto a national scale using district-wide proficiency rates, an approach that contrasts with our own.[1] Although some have argued that using NAEP as a basis for linking state accountability tests as we do here is both infeasible and inappropriate for high-stakes student-level reporting (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), our goal here is different. We do not attempt to estimate student-level scores, and do not intent the results to be used for high-stakes accountability. Rather, our goal is to estimate transformations that render aggregate test score distributions roughly comparable across districts in different states, so that the resulting district-level distributions can be used in aggregate-level research. Moreover, we treat the issue of feasibility empirically here, using a variety of validation checks to assess the extent to which our methods yield unbiased estimates of aggregate means and standard deviations.

Data

[1] Our data and methods are more comprehensive than a similar effort: Greene and McGee's *Global Report Card* (GRC, 2011; http://globalreportcard.org/). First, their district data are from 2004, 2005, and 2007, whereas ours are from 2009-2013. Second, we provide grade-by-year information, allowing for measures of progress. Third, instead of the statistical model we describe below (Reardon, Shear, Castellano, & Ho, 2016), which leverages information from three cut scores in each grade, the GRC uses only one cut score and aggregates across grades. This assumes that stringency is the same across grades and that district variances are equal. Fourth, our methods allow us to provide standard errors for our estimates. Fifth, we provide both direct and indirect validation checks for our linkages.

We use data corresponding to the first three rows of Table 1: state accountability test score data; NAEP data; and NWEA MAP data. Under the EDFacts Initiative (U.S. Department of Education, 2015), states provide frequencies of students in ordinal proficiency categories for each tested school, grade, and subject (mathematics and reading/ELA). The numbers of ordered proficiency categories vary by state, from 2 to 5. We use EdFacts data from 2009 to 2013, in grades 3-8, provided to us by the National Center for Education Statistics under a restricted data use license. These data have no suppression or minimum cell size. The terms of our data use agreement allow us to report individual district-level means and standard deviations so long as a) no cell is reported where the number of students tested was less than 20; and b) we add a very small amount of noise to the estimates.[2] We also use reliability estimates collected from state technical manuals and reports for these same years and grades, imputing when they are not reported.[3]

States and participating TUDA districts report average NAEP scores and their standard deviation in odd years, in grades 4 and 8, in reading and mathematics. In each state and TUDA district, these scores are based on an administration of the NAEP assessments to random samples of students in the relevant grades and years. We use years 2009, 2011, and 2013 as a basis for linking; we use additional odd years from 2003 through 2007 as part of the validation analyses. State and TUDA district means and standard deviations, as well as their standard errors, are available from the NAEP Data Explorer (U.S. Department of Education, n.d.). To account for NAEP initiatives to expand and standardize inclusion of English learners and students with disabilities over this time period, we rely on the Expanded Population Estimates of means and standard deviations provided by the National Center of Education Statistics (see Braun, Zhang

[2] We add random error $e_i \sim N(0, \omega_i^2/n_i)$, where ω_i^2 is the sampling variance of a parameter (e.g., a district's mean or standard deviation) and n_i is the number of students tested in the relevant cell.

[3] From 2009-2012, around 70% of 2,400 state(50)-grade(6)-subject(2)-year(4) reliability coefficients were available. Missing reliabilities were imputed as predicted values from a linear regression of reliability on state, grade, subject, and year. Reliabilities from 2013, which were not yet available when these data were gathered, were assumed to be the same as corresponding reliabilities from 2012. As Reardon and Ho (2015) show, reliabilities are almost always within a few hundredths of 0.90, so imputation errors are not likely to be consequential.

& Vezzu, 2008; McLaughlin, 2005; National Institute of Statistical Sciences, 2009)..[4]

Finally, we use data from the NWEA MAP test that overlap with the years, grades, and subjects available in the EdFacts data: 2009-2013, grades 3-8, in reading and mathematics. Student-level MAP test score data (scale scores) were provided to us through a restricted-use data sharing agreement with NWEA. Several thousand school districts chose to administer the MAP assessment in some or all years and grades that overlap with our EdFacts data. Participation in the NWEA MAP is generally binary in districts administering the MAP; that is, in participating districts, either very few students or essentially all students are assessed. We exclude cases in any district's grade, subject, and year, where the ratio of assessed students to enrolled students is lower than 0.9 or greater than 1.1. This eliminates districts with scattered classroom-level implementation as well as very small districts with accounting anomalies (roughly 10% of the districts using the NWEA MAP tests). After these exclusions, we estimate district means and standard deviations from student-level data reported on the continuous MAP scale.

Linking Methods

The first step in linking the state test scores to a common scale is to convert the coarsened proficiency count data available in the EdFacts data to district means and standard deviations expressed on a continuous within-state scale. We do this in each state, separately in each grade, year, and subject. For this, we use the methods described in detail by Reardon, Shear, Castellano, and Ho (2016). In brief, they demonstrate that a heteroskedastic probit (HETOP) model can be used to estimate group (district) test score means and standard deviations from coarsened data; the resulting estimates are generally unbiased and are only slightly less precise than estimates obtained from (uncoarsened) student-level scale score data in typical state and national educational testing contexts. We refer readers to their paper for technical specifics. Because most states do not report district-level means and standard deviations,

[4] Key estimates have correlations near unity to those from regular estimates, and our central substantive conclusions are unchanged when we use regular estimates in our analyses.

the ability to estimate these distributional parameters from coarsened proficiency category data is essential, given that such categorical data are much more readily available (e.g., EdFacts). Of course, if individual scale score data or district-level means and standard deviations were readily available, this step would be unnecessary.

Fitting the HETOP model to EdFacts data yields estimates of each district's mean test score, where the means are expressed relative to the state's student-level population mean and standard deviation within a given grade, year, and subject. We denote these estimated district means and standard deviations as $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$, respectively, for district d, year y, grade g, and subject b. The HETOP estimation also provides standard errors of these estimates, denoted $se\left(\hat{\mu}^{\text{state}}_{dygb}\right)$ and $se\left(\hat{\sigma}^{\text{state}}_{dygb}\right)$, respectively (Reardon, Shear, Castellano, & Ho, 2016).

The second step of the linking process, illustrated in Figure 1, is to estimate a linear transformation linking each state/year/grade/subject standardized scale (the scale of $\hat{\mu}^{\text{state}}_{dygb}$) to its corresponding NAEP distribution. Recall that we have estimates of NAEP means and standard deviations at the state (s) level, denoted $\hat{\mu}^{\text{naep}}_{sygb}$ and $\hat{\sigma}^{\text{naep}}_{sygb}$, respectively, as well as their standard errors. To obtain estimates of these parameters in grades (3, 5, 6, and 7) and years (2010 and 2012) in which NAEP was not administered, we interpolate and extrapolate linearly. First, within each NAEP-tested year, 2009, 2011, and 2013, we interpolate and extrapolate from grades 4 and 8 to grades 3, 5, 6, and 7. Next, for all grades 3-8, we interpolate between the NAEP-tested years to estimate parameters in 2010 and 2012. We illustrate this below for means, and we apply the same approach to standard deviations. Note that this is equivalent to interpolating between years first and then interpolating and extrapolating to grades.

$$\begin{aligned}
\hat{\mu}^{naep}_{sygb} &= \hat{\mu}^{naep}_{sy4b} + \tfrac{g-4}{4}\left(\hat{\mu}^{naep}_{sy8b} - \hat{\mu}^{naep}_{sy4b}\right) \\
\hat{\mu}^{naep}_{s2010gb} &= \frac{1}{2}\left(\hat{\mu}^{naep}_{s2009gb} + \hat{\mu}^{naep}_{s2011gb}\right) \\
\hat{\mu}^{naep}_{s2012gb} &= \frac{1}{2}\left(\hat{\mu}^{naep}_{s2011gb} + \hat{\mu}^{naep}_{s2013gb}\right).
\end{aligned} \tag{1}$$

We evaluate the viability of linking to reported interpolated NAEP grades and years explicitly in this paper.

As Figure 1 illustrates, we proceed under the assumption that NAEP and state test score means and variances should be the same. Because district test score moments are already expressed on a state scale with mean 0 and unit variance, the estimated mapping of the standardized test scale in state s, year y, grade g, and subject b to the NAEP scale is given by Equation (2) below, where $\hat{\rho}^{\text{state}}_{sygb}$ is the estimated reliability of the state test. Given $\hat{\mu}^{\text{state}}_{dygb}$, this mapping yields an estimate of the of the district average performance on the NAEP scale; denoted $\widehat{\hat{\mu}^{\text{naep}}_{dygb}}$. Given this mapping, the estimated standard deviation, on the NAEP scale, of scores in district d, year y, grade g, and subject b is given by Equation (3).

$$\widehat{\hat{\mu}^{\text{naep}}_{dygb}} = \hat{\mu}^{\text{naep}}_{sygb} + \frac{\hat{\mu}^{\text{state}}_{dygb}}{\sqrt{\hat{\rho}^{\text{state}}_{sygb}}} * \hat{\sigma}^{\text{naep}}_{sygb} \tag{2}$$

$$\widehat{\hat{\sigma}^{\text{naep}}_{dygb}} = \left[\frac{\left(\hat{\sigma}^{\text{state}}_{dygb}\right)^2 + \hat{\rho}^{\text{state}}_{sygb} - 1}{\hat{\rho}^{\text{state}}_{sygb}}\right]^{1/2} \cdot \hat{\sigma}^{\text{naep}}_{sygb} \tag{3}$$

The intuition behind Equation (2) is straightforward and illustrated in Figure 1: districts that belong to states with relatively high NAEP averages, $\hat{\mu}^{\text{naep}}_{sygb}$, should be placed higher on the NAEP scale. Within states, districts that are high or low relative to their state (positive and negative on the standardized state scale) should be relatively high or low on the NAEP scale in proportion to that state's NAEP standard deviation, $\hat{\sigma}^{\text{naep}}_{sygb}$.

The reliability term, $\hat{\rho}^{\text{state}}_{sygb}$, in Equations (2) and (3) is necessary to account for measurement error in state accountability test scores. Note that district means on the state scale, $\hat{\mu}^{\text{state}}_{dygb}$, are expressed in terms of standard deviation units of the state score distribution; thus, these standardized means are attenuated toward zero due to measurement error. They must be disattenuated before being mapped to the NAEP scale, given that NAEP scale accounts for measurement error due to item sampling. We disattenuate them by dividing the means by the square root of the state test score reliability estimate,

$\hat{\rho}_{sygb}^{\text{state}}$. The district standard deviations on the state scale, $\hat{\sigma}_{dygb}^{state}$, are biased toward 1 due to measurement error; we adjust them before linking them to the NAEP scale, as shown in Equation (3).

Treating the main terms in Equations (2) and (3) as independent random variables, we can derive the (squared) standard errors of the linked means and standard deviations:

$$var\left(\hat{\mu}_{dygb}^{\widehat{\text{naep}}}\right) = var\left(\hat{\mu}_{sygb}^{\text{naep}}\right) + \frac{var\left(\hat{\sigma}_{sygb}^{\text{naep}}\right)var\left(\hat{\mu}_{dygb}^{\text{state}}\right)}{\hat{\rho}_{sygb}^{\text{state}}} + \frac{\left(\hat{\sigma}_{sygb}^{\text{naep}}\right)^2 var\left(\hat{\mu}_{dygb}^{\text{state}}\right)}{\hat{\rho}_{sygb}^{\text{state}}} + \frac{\left(\hat{\mu}_{dygb}^{\text{state}}\right)^2 var\left(\hat{\sigma}_{sygb}^{\text{naep}}\right)}{\hat{\rho}_{sygb}^{\text{state}}} \quad (4)$$

$$var\left(\hat{\sigma}_{dygb}^{\widehat{\text{naep}}}\right) = z \cdot var\left(\hat{\sigma}_{dygb}^{\text{state}}\right)var\left(\hat{\sigma}_{sygb}^{\text{naep}}\right) + z \cdot var\left(\hat{\sigma}_{dygb}^{\text{state}}\right)\left(\hat{\sigma}_{sygb}^{\text{naep}}\right)^2 + var\left(\hat{\sigma}_{sygb}^{\text{naep}}\right)\left(\frac{\left(\hat{\sigma}_{dygb}^{\text{state}}\right)^2 + \hat{\rho}_{sygb}^{\text{state}} - 1}{\hat{\rho}_{sygb}^{\text{state}}}\right), \quad (5)$$

where

$$z = \frac{\left(\hat{\sigma}_{dygb}^{\text{state}}\right)^2}{\hat{\rho}_{sygb}^{\text{state}}\left(\left(\hat{\sigma}_{dygb}^{\text{state}}\right)^2 + \hat{\rho}_{sygb}^{\text{state}} - 1\right)}.$$

Validation Checks and Results

The linking method we use here, on its own, is based on the untested assumption that districts' distributions of scores on the state accountability tests have the same relationship to one another (i.e., the same relative means and standard deviations) as they would if the NAEP assessment were administered in lieu of the state test. Implicit in this assumption is that differences in the content, format, and testing conditions of the state and NAEP tests do not differ in ways that substantially affect aggregate relative distributions. This is, on its face, a strong assumption.

Rather than assert that this assumption is valid, we empirically assess it. We do this in several

ways. First, as illustrated in Figure 1, for the 20 districts[5] participating in the NAEP TUDA assessments, we compare $\widehat{\hat{\mu}_{dygb}^{\text{naep}}}$—the estimated district mean based on our linking method—to $\hat{\mu}_{dygb}^{\text{naep}}$—the mean of NAEP TUDA scores from the district. This provides a direct validation of the linking method, since the TUDA scores are in the metric the linking method attempts to recover but are not themselves used in any way in the linking process. Second, we assess whether within-district differences in linked scores across grades and cohorts correspond to those differences observed in the TUDA data. That is, we assess whether the linking provides accurate measures of changes in scores across grades and cohorts of students, in addition to providing accurate means in a given year. Third, we assess the correlation of our linked district estimates with district mean scores on the NWEA MAP tests. This provides an assessment of the correlation across a larger sample of districts but uses a different test, so it does not provide direct comparability with the NAEP scale that is the target of our linking. Fourth, we conduct a set of validation exercises designed to assess the validity of the interpolation of the NAEP scores in non-NAEP years and grades. For all of these analyses, we present evidence regarding the district means; corresponding results for the standard deviations are in the appendices.

Validation Check 1: Recovery of TUDA means

The bottom of Figure 1 illustrates the first validation check. The NAEP TUDA data provide means and standard deviations on the actual "naep" scale, $\hat{\mu}_{dygb}^{\text{naep}}$ and $\hat{\sigma}_{dygb}^{\text{naep}}$ for 17 large urban districts in 2009 and 20 in 2011 and 2013.[6] For these particular large districts, we can compare the NAEP means and standard deviations to their linked means and standard deviations. For each district, we obtain a

[5] This excludes Washington, DC, which does not have a respective state distribution for validation.
[6] In 2009, the 17 districts are Atlanta, Austin, Baltimore, Boston, Charlotte, Chicago, Cleveland, Detroit, Fresno, Houston, Jefferson County, Los Angeles, Miami, Milwaukee, New York City, Philadelphia, and San Diego. Albuquerque, Dallas, and Hillsborough County joined in 2011 and 2013. Washington, DC is not included for validation, as it has no associated state for linking. California districts do not have a common Grade 8 state mathematics assessment, so the three California districts lack a linked district mean for Grade 8 mathematics.

discrepancy, $\widehat{\hat{\mu}_{dygb}^{\text{naep}}} - \hat{\mu}_{dygb}^{\text{naep}}$. We report the average of these discrepancies as the bias, and we report the square root of the average squared discrepancies as the Root Mean Squared Error (RMSE). We also report the correlation between the two. Because of imprecision in both the NAEP TUDA and linked estimates, the RMSE will be inflated and the correlation will be attenuated as measures of recovery. We report disattenuated correlations that account for imprecision in both $\widehat{\hat{\mu}_{dygb}^{\text{naep}}}$ and $\hat{\mu}_{dygb}^{\text{naep}}$.

Table 2 reports the results of these analyses in each subject, grade, and year in which we have TUDA estimates. On average, the linked estimates overestimate actual NAEP TUDA means by roughly 2 points on the NAEP scale, or around .06 of a standard deviation unit, assuming the original NAEP scale standard deviation of 35 (NAEP standard deviations vary from roughly 30 to 40 across subjects, years, and grades). The bias is slightly greater in earlier years and in mathematics.

Table 2 here

This positive bias indicates that the average scores of students in the TUDA districts are systematically higher in the statewide distribution of scores on the state accountability tests than on the NAEP test. This leads to a higher-than-expected NAEP mapping. Table 2 also shows that the average correlation between the observed linked and actual TUDA means, within a grade, year, and subject, is high (0.94). The average estimated true correlation (disattenuated to account for the imprecision in the observed means) is 0.95. Figure 2 shows scatterplots of the estimated linked means versus the observed TUDA means, separately for grades and subjects, with the identity lines displayed as a reference.

Note that under a linear linking such as Equation 2, our definition of bias implies that weighted average bias, among all districts within each state, and across all states, is 0 by design. The bias in Table 2 is not 0 because Table 2 summarizes the bias for only the subset of districts for which we have NAEP scores.

We review here four of the number of possible explanations for discrepancies between a district's average scores on the state accountability test and on the NAEP assessments. First, the

population of students assessed in the two instances may differ. For example, a positive discrepancy may result if the target district excluded low scoring students from state tests but not from NAEP. If this differential exclusion were greater in the target district, on average, than in other districts in the state, the target district would appear higher in the state test score distribution than it would in the NAEP score distribution, leading to a positive discrepancy between the district's linked mean score and its NAEP mean scores. Likewise, a positive discrepancy would result if the NAEP assessments excluded high scoring students more in the TUDA assessment than in the statewide assessment, or if there were differential exclusion of high-scoring students in other districts on the state test relative to the target district and no differential exclusion on NAEP. In other words, the discrepancies might result from a target district's scores being biased upward on the state test or downward on the NAEP assessment relative to other districts in the state, and/or from other districts' scores being biased downward on the state test or upward on the NAEP assessment relative to the target district.

Second, the discrepancies may result from differential content in NAEP and state tests. If a district's position in the state distribution of skills/knowledge measured by the state test does not match its position in the statewide distribution of skills measured by the NAEP assessment, the linked scores will not match those on NAEP. The systematic positive discrepancies in Table 2 and Figure 2 may indicate that students in the TUDA districts have disproportionately higher true skills in the content areas measured by their state tests than the NAEP assessments relative to other districts in the states. In other words, if large districts are better than other districts in their states at teaching their students the specific content measured by state tests, relative to their effectiveness in teaching the skills measured by NAEP, we would see a pattern of positive discrepancies like that in Table 2 and Figure 2.

Third, relatedly, students in the districts with a positive discrepancy may have relatively high motivation for state tests over NAEP, compared to other districts. Fourth, the bias evident in Table 2 and Figure 2 may indicate relative inflation or outright cheating. For example, some of the largest positive

discrepancies among the 20 TUDA districts illustrated in Figure 2 are in Atlanta in 2009, where there was systematic cheating on the state test in 2009. The discrepancies in the Atlanta estimates are substantially smaller (commensurate with other large districts) in 2011 and 2013, after the cheating had been discovered. In this way, we see that many possible sources of bias in the linking are sources of bias with district scores on the state test itself, rather than problems with the linking per se.

Validation Check 2: Association with NWEA MAP means

The NWEA MAP test is administered in thousands of school districts across the country. Because the MAP tests are scored on the same scale nationwide, district average MAP scores can serve as a second audit test against which we can compare the linked scores. As noted previously, in most such districts, the number of student test scores is very close to the district's enrollment in the same subject, grade, and year. For these districts, we estimate means and standard deviations on the scale of the MAP test, which we designate "$\mathbf{map}$". The scale differs from that of NAEP, so absolute discrepancies are not interpretable. However, strong correlations between linked district means and standard deviations and those on MAP represent convergent evidence that the linking is appropriate. We calculate disattenuated correlations between observed MAP and linked means before and after the linkage in Equations 2 and 3. For means:

$$\text{Before: } Corr\left(\hat{\mu}^{\text{state}}_{dygb}, \hat{\mu}^{\text{map}}_{dygb}\right).$$

$$\text{After: } Corr\left(\widehat{\hat{\mu}^{\text{naep}}_{dygb}}, \hat{\mu}^{\text{map}}_{dygb}\right).$$

(6)

Table 3 shows that correlations between "post linked" district means and MAP district means are 0.93 when adjusting for imprecision due to measurement error. These "post-link" correlations are increases from "pre-link" correlations of 0.87. Figure 3 shows a bubble plot of district MAP scores on linked scores for Grade 4 mathematics in 2009, 2011, and 2013, as an illustration of the data underlying

these correlations. Note that the points plotted in Figure 3 are the estimated means, and so are subject to measurement error; the true correlation is higher than that illustrated here.[7]

Table 3 here

Validation Check 3: Association of between-grade and –cohort trends

An additional assessment of the extent to which the linked state district means match the corresponding NAEP or NWEA district means compares not just the means in a given grade and year, but compares the within-district differences in means across grades and years. If the discrepancies evident in Figure 2 are consistent across years and grades within a district, then the linked state estimates will provide accurate measures of the within-district trends across time and grade, even when there is a small bias in in the average means.

To assess the accuracy of the between-grade and -year differences in linked mean scores, we use data from the grades and years in which we have both linked means and corresponding means from NAEP. We do not use the NAEP data from interpolated years and grades in this model. We fit a similar model using the NWEA data as the audit test; here we use data from all grades and years for which we have both NWEA data and linked estimates (including years in which the linking is based on NAEP interpolations). For both models, we fit precision-weighted random coefficients models of this form:

$$\hat{\mu}_{dygbi} = \alpha_{0dygb}(LINK) + \alpha_{1dygb}(TARGET) + e_{dygbi}$$

$$\alpha_{0dygb} = \beta_{00d} + \beta_{01d}(year_{dygb} - 2011) + \beta_{02d}(grade_{dygb} - 6) + u_{0dygb}$$

$$\alpha_{1dygb} = \beta_{10d} + \beta_{11d}(year_{dygb} - 2011) + \beta_{12d}(grade_{dygb} - 6) + u_{1dygb}$$

$$\beta_{00d} = \gamma_{00} + v_{00d}$$

$$\beta_{01d} = \gamma_{01} + v_{01d}$$

$$\beta_{02d} = \gamma_{02} + v_{02d}$$

[7] The observed (attenuated) correlations are generally .05 to .09 points lower than their disattenuated counterparts.

$$\beta_{10d} = \gamma_{10} + v_{10d}$$

$$\beta_{11d} = \gamma_{11} + v_{11d}$$

$$\beta_{12d} = \gamma_{12} + v_{12d}$$

$$e_{dygbi} \sim N\left(0, \omega^2_{dygbi}\right);\ \boldsymbol{u}_{dygb} \sim MVN(0, \boldsymbol{\sigma}^2);\ ;\ \boldsymbol{v}_d \sim MVN(0, \boldsymbol{\tau}^2),$$

(7)

where i indexes source (linked or NAEP/NWEA test) and ω^2_{dygbi} is the sampling variance of $\hat{\mu}_{dydbi}$. The vector $\Gamma = \{\gamma_{00}, \ldots, \gamma_{12}\}$ contains the average intercepts, year slopes, and grade slopes on each of the two tests. When we use NAEP as the target test, differences between the corresponding elements of Γ indicate average bias (i.e., the difference between γ_{00} and γ_{10} indicates the average deviation of the linked means and the NAEP TUDA means, net of district-specific grade and year trends. Unlike Table 2 above, where we estimated bias separately for each year and grade, the bias here is estimated by pooling over all years and grades of TUDA data. If the linking were perfect, we expect this to be 0.

The matrix of random coefficients $\boldsymbol{\tau}^2$ includes, on the diagonal, the between-district variances of the average district means and their grade and year trends; the off-diagonal elements are their covariances. From $\boldsymbol{\tau}^2$ we can compute the correlation between the within-district differences in mean scores between grades and years. The correlation $corr(v_{01d}, v_{11d})$, for example, describes the correlation between the temporal trend in district NAEP scores and the trend in the linked scores. Likewise the correlation $corr(v_{02d}, v_{12d})$ describes the correlation between the grade 4-8 differences in district NAEP scores and the corresponding difference in the linked scores. Finally, the correlation $corr(v_{00d}, v_{10d})$ describes the correlation between the NAEP and linked intercepts in the model—that is, the correlation between average linked and TUDA mean scores. This correlation differs from that shown in Table 2 above because the former estimates the correlation separately for each grade and year; Model (7) estimates the correlation from a model in which all years and grades are pooled.

Table 4 shows the results from fitting this model (separately by subject). When comparing the

linked estimates to the NAEP TUDA estimates, several patterns are evident. First, the estimated correlation of the TUDA and linked intercepts is 0.98 (for both math and reading) and the bias in the means (the difference in the estimated intercepts in Table 4) is small and not statistically significant. The linked ELA means are, on average 1.1 points higher (s.e. of the difference is 3.0; n.s) than the TUDA means; and the linked math means are, on average, 2.5 points higher (s.e. of the difference is 3.3, n.s.) than the TUDA means. These are, not surprisingly, similar to the average bias estimated from each year and grade separately and shown in Table 2. Second, the estimated average linked and TUDA grade slopes are nearly identical to one another; this is true in both math and ELA. The estimated correlation of the TUDA and linked grade slopes is 0.85 for ELA and 0.99 for math. The reliability of the grade differences of the linked estimates is 0.76 in ELA and 0.74 in math. This indicates that the linked estimates provide unbiased estimates of the differences within districts across grades, and that these estimates are precise enough to carry meaningful information about between-grade differences. Third, there is little or no variation in the year trends in the TUDA districts; for both math and reading, the estimated variation of year trends is small and not statistically significant. As a result, neither the TUDA nor the linked estimates provide estimates of trends across grades that are sufficiently reliable to be useful (in models not shown, we estimate the reliabilities of the TUDA year slopes to be 0.28 and 0.53 and of the linked year slopes to be 0.45 and 0.72 in ELA and math, respectively). As a result, we dropped the random effects on the year trends and do not report an estimate of the correlation of the year trends for the TUDA data.

Table 4 here

Validation Check 4: Recovery of estimates under interpolation within years

Although we cannot assess recovery of linkages in interpolated grades with only grades 4 and 8, we can check recovery for an interpolated year, specifically, 2011, between 2009 and 2013. By pretending that we do not have 2011 data, we can assess performance of our interpolation approach by

comparing linked estimates to actual 2011 TUDA results. For each of the TUDAs that participated in both 2009 and 2013, we interpolate, for example,

$$\hat{\mu}^{naep'}_{s2011gb} = \frac{1}{2}\left(\hat{\mu}^{naep}_{s2009gt} + \hat{\mu}^{naep}_{s2013gt}\right)$$

$$\hat{\sigma}^{naep'}_{s2011gb} = \frac{1}{2}\left(\hat{\sigma}^{naep}_{s2009gt} + \hat{\sigma}^{naep}_{s2013gt}\right)$$

(8)

Applying Equations 2-5, we obtain estimates, for example, $\hat{\mu}^{\widehat{\text{naep}}'}_{d2011gb}$, and we compare these to actual TUDA estimates from 2011. We estimate discrepancies as $\hat{\mu}^{\widehat{\text{naep}}'}_{d2011gb} - \hat{\mu}^{\text{naep}}_{d2011gb}$. Table 5 shows results in the same format as Table 2. We note that the average RMSE of 3.9 and bias of 1.5 in Table 5 are approximately the same as the average RMSE of 3.9 and bias of 1.6 shown for 2011 in Table 2. A different perspective on this same finding is that interpolated 2011 means very accurately match the reported means. Note that the interpolations we actually use interpolate between observed scores that are only two years apart, rather than four years apart as in the exercise here. The two-year interpolations are almost certainly more accurate than the four-year interpolation (which itself is accurate enough to show no degradation in our recovery of estimated means). We conclude that the between-year interpolation of state NAEP scores adds no appreciable error to the linked estimates.

Table 5 here

Validation Check 5: Association with NWEA MAP means across degrees of interpolation

We further investigate the viability of interpolation by comparing correlations of linked district estimates with MAP scores for interpolated and uninterpolated scores. Some grade-year combinations need no interpolation, others are singly interpolated, and others are doubly interpolated.

Table 6 shows that, on average, precision-adjusted correlations between linked NAEP means and MAP means are almost identical across different degrees of interpolation, around 0.93. This lends additional evidence that interpolation adds negligible aggregate error to recovery.

Table 6 here

Validation Check 6: Reliability of interpolated means

The recovery of means under interpolation is a testament to the reliability of state NAEP means and standard deviations across grades and over time. We can assess this in part by standardizing NAEP scores within years, grades, and subjects, and evaluating the stability of state means and standard deviations across grades and over years. For these analyses we can use data from NAEP grades 4 and 8 and across the years 2005, 2007, 2009, 2011, and 2013. If the relative position of state means and standard deviations on NAEP are stable across grades 4 and 8, for example, this builds a case for extending the linkage across other grades. Following reliability and generalizability theory (Haertel, 2006), we fit the crossed random effects model:

$$\mu_{syg} = \mu + \nu_s + \nu_{sy} + \nu_{sg} + \nu_{syg,e}.$$

And we estimate the following intraclass correlation parameter indicating the reliability when averaging over one or more years, n_y, or grades, n_g.

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{sy}^2}{n_y} + \frac{\sigma_{sg}^2}{n_g} + \frac{\sigma_{syg,e}^2}{n_y n_g}}$$

Table 7 shows that the reliability of singly and doubly interpolated means is very high, from around 0.90 to 0.95. This suggests that states' relative means change little between 4th and 8th grade and across years. This helps to explain the good recovery results under Validation Checks 4 and 5.

Table 7 here

Scaling

The linked estimates of districts' mean test scores across grades, years, and subjects are expressed on the NAEP math and ELA scales. It may be useful to convert the NAEP scale to a metric more

useful for analysis and interpretation. Here we describe methods of transforming the NAEP scale to three different metrics: a scale measured in grade-, year-, and subject-specific national population standard deviation units; a scale measured in grade- and subject-specific standard deviation units of the national population score distribution in a given cohort; and a scale measured in units of average grade-level differences in scores.

First, we standardize the estimates relative to the national distribution of NAEP scores in each year, grade, and subject. That is, we compute:

$$\hat{\mu}_{dygb}^{\widehat{\mathrm{n*}}} = \frac{\hat{\mu}_{dygb}^{\widehat{\mathrm{naep}}} - \hat{\mu}_{ygb}^{\mathrm{naep}}}{\hat{\sigma}_{ygb}^{\mathrm{naep}}} \tag{9}$$

and

$$\hat{\sigma}_{dygb}^{\widehat{\mathrm{n*}}} = \frac{\hat{\sigma}_{dygb}^{\widehat{\mathrm{naep}}}}{\hat{\sigma}_{ygb}^{\mathrm{naep}}}, \tag{10}$$

where $\hat{\mu}_{ygb}^{\mathrm{naep}}$ and $\hat{\sigma}_{ygb}^{\mathrm{naep}}$ are the national mean and standard deviation of NAEP scores in a given year, grade, and subject. We interpolate as above (see Equation 1) to estimate these in 2010 and 2012 and in grades 3, 5, 6, and 7. The problem with this method of standardization is that it destroys information about real changes over time in the means and standard deviations of scores that is contained in the changing means and standard deviations of the national NAEP score distribution.

An alternate is to standardize the scores to a common mean and standard deviation across years (but different across grades). We do this as follows. We first estimate the average (linear) within-cohort change in subject b test scores per grade, denoted γ_b, by using the published estimates of the national Main NAEP means and standard deviations in grades 4 and 8 (indexed by g) in 2009, 2011, and 2013 (indexed by y) to fit the models

$$\hat{\mu}_{ygb}^{\text{naep}} = \alpha_{\mu b} + \beta_{\mu b}(y - g - 2005.5) + \gamma_{\mu b}(g - 5.5) + e_{\mu}.$$

$$\hat{\sigma}_{ygb}^{\text{naep}} = \alpha_{\sigma b} + \beta_{\sigma b}(y - g - 2005.5) + \gamma_{\sigma b}(g - 5.5) + e_{\sigma}.$$

(11)

We fit these models separately for math and reading (indexed by b).

Now $\alpha_{\mu b}$ and $\alpha_{\sigma b}$ are estimates of the interpolated mean and standard deviation, respectively, of the NAEP score distribution in subject b in 2011 in grade 5.5 (the middle of our data). We estimate the mean and standard deviation of the national distribution in grade g for this cohort of students as

$$\hat{\mu}_{gb}^{naep} = \hat{\alpha}_{\mu b} + \hat{\gamma}_{\mu b}(g - 5.5)$$

$$\hat{\sigma}_{gb}^{naep} = \hat{\alpha}_{\sigma b} + \hat{\gamma}_{\sigma b}(g - 5.5).$$

(12)

We then use these estimated national means and standard deviations to standardize the linked estimates relative to the 2005 cohort's national distribution of NAEP scores in each grade, and subject. That is, we compute:

$$\hat{\mu}_{dygb}^{\widehat{c*}} = \frac{\hat{\mu}_{dygb}^{\widehat{\text{naep}}} - \hat{\mu}_{gb}^{\text{naep}}}{\hat{\sigma}_{gb}^{\text{naep}}}$$

(13)

and

$$\hat{\sigma}_{dygb}^{\widehat{c*}} = \frac{\hat{\sigma}_{dygb}^{\widehat{\text{naep}}}}{\hat{\sigma}_{gb}^{\text{naep}}}.$$

(14)

The district test score distributions are now standardized using the estimated grade-specific national distribution of scores from a common cohort of students.

The third way of standardizing the estimates is to convert them to (approximate) units of average between-grade score differences. To do this, we use the estimates of $\alpha_{\mu b}$ and $\gamma_{\mu b}$ from Equation (11)

above. $\hat{\gamma}_b$ is an estimate of the average within-cohort change in NAEP test scores per grade in the US population of students; $\hat{\alpha}_b$ is an estimate of the average NAEP score in the middle year and grade of our data (in grade 5.5 in 2011). We use $\hat{\alpha}_b$ and $\hat{\gamma}_b$ to standardize the district-year-grade-subject estimates

$$\hat{\mu}_{dygb}^{\widehat{g*}} = \frac{\hat{\mu}_{dygb}^{\widehat{naep}} - \hat{\alpha}_{\mu b}}{\hat{\gamma}_{\mu b}} \tag{15}$$

and

$$\hat{\sigma}_{dygb}^{\widehat{g*}} = \frac{\hat{\sigma}_{dygb}^{\widehat{naep}}}{\hat{\gamma}_{\mu b}}. \tag{16}$$

Now $\hat{\mu}_{dygb}^{\widehat{g*}}$ is the estimated average national grade-equivalent of students in district d, year y, grade g, and subject b. So if $\hat{\mu}_{dy4b}^{\widehat{g*}} = 5$, students in district d, year y are one grade level above the national average in subject t in 4th grade.

The three methods of standardization have different interpretations. The first (denoted $\hat{n}^*$) expresses districts' score distributions in units of grade-, year-, and subject-specific national population standard deviations. This method does not require that the NAEP scale is vertically linked across grades or is common across years. Its drawback is that it does not provide information on absolute changes in districts' score distributions over time or across grades. Given that the NAEP scale is designed to be stable over time within a grade and subject, the within- year standardization destroys useful information.

The second method of standardization (denoted $\hat{c}^*$) expresses districts' score distributions in units of a given cohort's grade-specific national standard deviation units. This scale retains information about absolute changes over time, but does so by relying on the stability of the NAEP scale over time and on the linear interpolation of NAEP distributions over time. Neither of those assumptions is problematic: NAEP is designed to have a stable scale over time, and the interpolation for 2010 and 2012 is very

reliable, as our analyses above show. This scale describes relative changes in districts' scores across grades, but does not provide information about absolute changes across grades, because the scale is standardized within each grade.

Finally, the third method of standardization (denoted $\hat{g}^*$) expresses districts' score distributions in units that correspond to national grade-level averages and differences. On this scale, a one-unit difference corresponds the national average within-cohort difference in scores between students in adjacent grades. The scale is set so that a value of 4, for example, corresponds to the average NAEP score among 4th graders in the middle cohort of our data; a value of 8 corresponds to the average NAEP score among 8th graders in that same cohort. This metric contains information on both absolute changes across grades and over time, but does so by relying on the linear interpolation of NAEP score means and standard deviations in grades other than 4 and 8 and years 2010 and 2012, and on the assumption that the NAEP scale is stable over time and vertically linked across grades. This scale is more readily interpretable, particularly to non-technical audiences, but may not be preferable for analyses where the vertical linking across grades and the linear interpolation assumptions are not required or defensible.

Discussion

A nationwide district-level dataset of test score means and standard deviations is a valuable tool for descriptive and causal analysis of academic achievement if and only if it is valid for its intended research purposes. We use a range of validation approaches to demonstrate that test score distributions on state standardized tests can be transformed to a common national NAEP-linked scale in a way that yields district-level distributions that correspond well—but not perfectly—to the relative performance of students in different districts on the NAEP and MAP assessments. The correlation of district-level mean scores on the NAEP-linked scale with scores on the NAEP TUDA and NWEA MAP assessments is generally high (averaging 0.95 and 0.93 across grades, years, and subjects, respectively). Nonetheless, we find

some evidence that NAEP-linked estimates include some small, but systematically positive, bias in large urban districts (roughly +0.06 standard deviations, on average). This implies a corresponding small downward bias for some other districts in the same states.

Are these discrepancies a threat to the validity of the linked estimates of district means? The answer depends on how the estimates will be used. Given the evidence of imperfect correlation and small bias, the linked estimates should not be used to compare or rank school districts' performance when the estimated means are close and when the districts are in different states (within-state comparisons do not depend on the linking procedure, so are immune to bias that arises from the linking methods).

The linked estimates are, however, clearly accurate enough to be used to investigate broad patterns of the relationship between average test performance and local community or schooling conditions, both within and between-states. The validation exercises suggest that the linked estimates can be used to examine variation among districts and across grades within districts. It is unclear whether the estimates provide unbiased estimates of trends over time, given that there is little or no variation in the NAEP TUDA districts' trends over time against which to benchmark the linked trend estimates.

Perhaps the most appropriate interpretation of the linked estimates is that they are the result of a set of monotonic transformations of districts' score distributions on state tests: they are state score distributions with NAEP-based adjustments, with credit given for being in a state with relatively high NAEP performance and, for districts within the states, greater discrimination among districts when a state's NAEP standard deviation is high. The resulting score distributions are useful to the extent districts' state test score distributions rank districts similarly as they would be ranked on the NAEP assessment. Because the testing conditions, purpose, motivation, and content of NAEP and state tests differ, these rankings, could we observe them, would differ. But our validation checks suggest that they would be more similar than different. This is evident in the high correspondence of the linked and NAEP TUDA estimates and of

the linked and NWEA MAP estimates. This suggest that our estimated NAEP-linked district test score means, which is unprecedented in its scope and geographical detail, may be very useful in empirical research describing and analyzing national variation in local academic performance.

References

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). *Mapping state proficiency standards onto NAEP scales: Results from the 2013 NAEP reading and mathematics assessments* (NCES 2015-046). U.S. Department of Education, Washington, DC: National Center for Education Statistics.

Braun, H., Zhang, J., and Vezzu, S. (2008). Evaluating the Effectiveness of a Full-Population Estimation Method. Educational Testing Service Research Report RR-08-18. http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02104.x/epdf

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M.W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17,* 267-321.

McLaughlin D. (2005). Properties of NAEP Full Population Estimates. Unpublished report, American Institutes for Research. http://www.schooldata.org/Portals/0/uploads/reports/NSA_T1.5_FPE_Report_090205.pdf

National Institute of Statistical Sciences (2009). NISS/NESSI Task Force on Full Population Estimates for NAEP. Technical Report #172. http://www.niss.org/sites/default/files/technical_reports/tr172.pdf

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2016). *Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data.* Retrieved from https://cepa.stanford.edu/sites/default/files/wp16-02-v201601.pdf

Reardon, S. F., Yun, J. T., & Eitle, T. M. (1999). *The changing context of school segregation: Measurement and evidence of multi-racial metropolitan area school segregation, 1989-1995*. Paper presented at the annual meeting of the American Educational Research Association. Montreal, Canada.

Stroub, K. J., & Richards, M. P. (2013). From resegregation to reintegration: Trends in the racial/ethnic segregation of metropolitan public schools, 1993–2009. *American Educational Research Journal, 50*, 497-531.

U.S. Department of Education. (2015). *EDFacts Submission System User Guide V11.2* (SY 2014-2015). Washington, DC: EDFacts. Retrieved from http://www.ed.gov/edfacts

U.S. Department of Education (n.d.), *NAEP Data Explorer*, Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences.

Whitehurst, G. J., Chingos M. M., & Gallaher, M. R. (2013). *Do School Districts Matter?* Washington, DC: Brookings Institution.

Table 1: Assessment programs that allow cross-state comparisons of state and district educational achievement.

Assessment	Scope	Frequency	Grades (of 3-8)	Maximum Comparisons as of 2013	
				State-to-State (out of 50)	Districts Across States (out of ~13500)
NAEP	National	Odd years	4, 8	50	21
NECAP*	4-State	2005-2014	3-8	4	984
NWEA MAP	National	1-3 / year	3-8	0	1000-2000
SBAC/PARCC	Consortium	Annual	3-8	0 (15/7 anticipated)	0 (will vary)
Our Linking	**National**	**Annual**	**3-8**	**50**	**~13500**

Note: NAEP = National Assessment of Educational Progress; NECAP = New England Common Assessment Program; NWEA MAP = Northwest Evaluation Association Measures of Academic Progress; SBAC/PARCC = Smarter Balanced Assessment Consortium/Partnership for Assessment of Readiness for College and Careers. * Defunct; **Anticipated

Table 2: Recovery of NAEP TUDA means following state-level linkage of state test score distributions to the NAEP scale.

			Recovery			Relationship	
Subject	Grade	Year	n	RMSE	Bias	Corr.	Adj. Corr.
Reading	4	2009	17	3.93	2.10	0.94	0.96
		2011	20	4.08	1.26	0.94	0.96
		2013	20	3.09	0.08	0.96	0.98
	8	2009	17	3.21	1.09	0.91	0.95
		2011	20	2.52	0.53	0.96	0.98
		2013	20	3.87	1.63	0.91	0.93
Math	4	2009	17	5.74	4.21	0.92	0.93
		2011	20	5.06	2.66	0.93	0.94
		2013	20	3.73	1.45	0.94	0.95
	8	2009	14	4.90	3.55	0.95	0.95
		2011	17	3.89	2.06	0.96	0.96
		2013	17	4.74	1.81	0.93	0.94
Average		2009	65	4.45	2.74	0.93	0.95
		2011	77	3.89	1.63	0.95	0.96
		2013	77	3.86	1.24	0.94	0.95
		All Years	219	4.06	1.87	0.94	0.95

Note: Using NAEP Expanded Population Estimates. Adjusted correlations account for imprecision in linked and target estimates.

Table 3: Precision-adjusted correlations of linked estimates with NWEA MAP district means before and after state-level linkage of state test score distributions to the NAEP scale.

				Precision-Adjusted Correlations	
Subject	Grade	Year	n	Pre-Link	Post-Link
Reading	4	2009	1134	0.90	0.95
		2011	1476	0.88	0.93
		2013	1820	0.92	0.95
	8	2009	946	0.85	0.91
		2011	1276	0.87	0.91
		2013	1584	0.89	0.92
Math	4	2009	1123	0.86	0.93
		2011	1467	0.83	0.90
		2013	1831	0.87	0.93
	8	2009	959	0.84	0.93
		2011	1287	0.86	0.93
		2013	1544	0.88	0.95
Average		2009	4162	0.86	0.93
		2011	5506	0.86	0.92
		2013	6779	0.89	0.94
		All Years	16447	0.87	0.93

Note: Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Only includes districts with >90% of enrollment reporting scores on NWEA MAP.

Table 4. Pooled recovery estimates for linked district means of TUDA (left) or NWEA (right) counterparts, by subject.

	Linked - TUDA Comparison				Linked - NWEA Comparison			
	ELA		Math		ELA		Math	
Intercept- TUDA/NWEA	227.414	***	248.13	***	214.456	***	223.946	***
	(2.195)		(2.485)		(0.106)		(0.124)	
Grade * TUDA/NWEA	10.843	***	9.674	***	4.551	***	6.152	***
	(0.222)		(0.169)		(0.014)		(0.021)	
Year * TUDA/NWEA	1.033	***	0.904	***	0.162	***	0.254	***
	(0.103)		(0.110)		(0.017)		(0.020)	
Intercept - Linked	228.527	***	250.596	***	241.393	***	262.055	***
	(2.004)		(2.116)		(0.234)		(0.231)	
Grade * Linked	10.809	***	9.594	***	11.388	***	10.722	***
	(0.273)		(0.289)		(0.033)		(0.036)	
Year * Linked	0.907	***	0.467	*	0.703	***	0.604	***
	(0.170)		(0.182)		(0.032)		(0.034)	
L2 Variance- TUDA/NWEA	0.83		1.24		2.222		3.032	
L2 Variance- Linked	2.51		2.66		2.986		3.710	
L3 Variance Intercept - TUDA/NWEA	9.79	*	11.07	*	5.588	*	6.545	*
L3 Variance Intercept - Linked	8.88	*	9.34	*	12.420	*	12.260	*
L3 Variance Year - TUDA/NWEA					0.600	*	0.727	*
L3 Variance Year - Linked					1.162	*	1.285	*
L3 Variance Grade - TUDA/NWEA	0.93	*	0.60	*	0.601	*	0.925	*
L3 Variance Grade - Linked	1.06	*	1.05	*	1.448	*	1.649	*
Correlation - Intercepts	0.98		0.98		0.841		0.857	
Correlation - Grade Slopes	0.85		0.99		0.577		0.616	
Correlation - Year Slopes					0.394		0.600	
Reliability Intercept - TUDA/NWEA	1.00		1.00		0.925		0.920	
Reliability Grade- TUDA/NWEA	0.87		0.72		0.653		0.719	
Reliability Year- TUDA/NWEA					0.524		0.504	
Reliability Intercept - Linked	0.98		0.98		0.936		0.938	
Reliability Grade- Linked	0.76		0.74		0.719		0.762	
Reliability Year- Linked					0.522		0.558	
N - Observations	228		210		104958		103966	
N - Districts	20		20		2947		2945	

Table 5. Recovery of reported 2011 NAEP TUDA means following state-level linkage of state test score distributions to a NAEP scale interpolated between 2009 and 2013.

			Recovery			Relationship	
Subject	Grade	Year	n	RMSE	Bias	Corr.	Adj. Corr.
Reading	4	2011	20	4.13	0.82	0.94	0.95
	8	2011	20	2.64	1.28	0.96	0.99
Math	4	2011	20	4.82	2.19	0.92	0.94
	8	2011	17	4.00	1.70	0.95	0.96
		Average	77	3.90	1.50	0.94	0.96

Note: Using NAEP Expanded Population Estimates. Adjusted correlations account for imprecision in linked and target estimates.

Table 6: Precision-adjusted correlations between NWEA MAP district means and NAEP-linked estimates.

Subject	Grade	2009	2010	2011	2012	2013
Reading	3	0.95	0.94	0.94	0.94	0.94
	4	0.95	0.95	0.93	0.95	0.95
	5	0.95	0.94	0.93	0.94	0.94
	6	0.93	0.94	0.93	0.94	0.94
	7	0.92	0.93	0.92	0.93	0.93
	8	0.91	0.91	0.91	0.92	0.92
Math	3	0.91	0.90	0.91	0.91	0.91
	4	0.93	0.92	0.90	0.92	0.93
	5	0.92	0.91	0.92	0.92	0.93
	6	0.93	0.93	0.94	0.94	0.95
	7	0.95	0.95	0.95	0.95	0.95
	8	0.93	0.94	0.93	0.94	0.95

No interpolation	0.929	Reading	0.93
Single interpolation	0.932	Math	0.93
Double interpolation	0.931		

Note. Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Only includes districts with >90% of enrollment reporting scores on NWEA MAP.

Table 7: Reliabilities of average NAEP state means and standard deviations over grades (4, 8) and years (2003, 2005, 2007, 2011, 2013)

	1 grade-year	2 grades	2 years	2 grades, 2 years
Reading	0.86	0.91	0.89	0.94
Math	0.89	0.92	0.92	0.95

Figure 1. Illustration of linear linking method

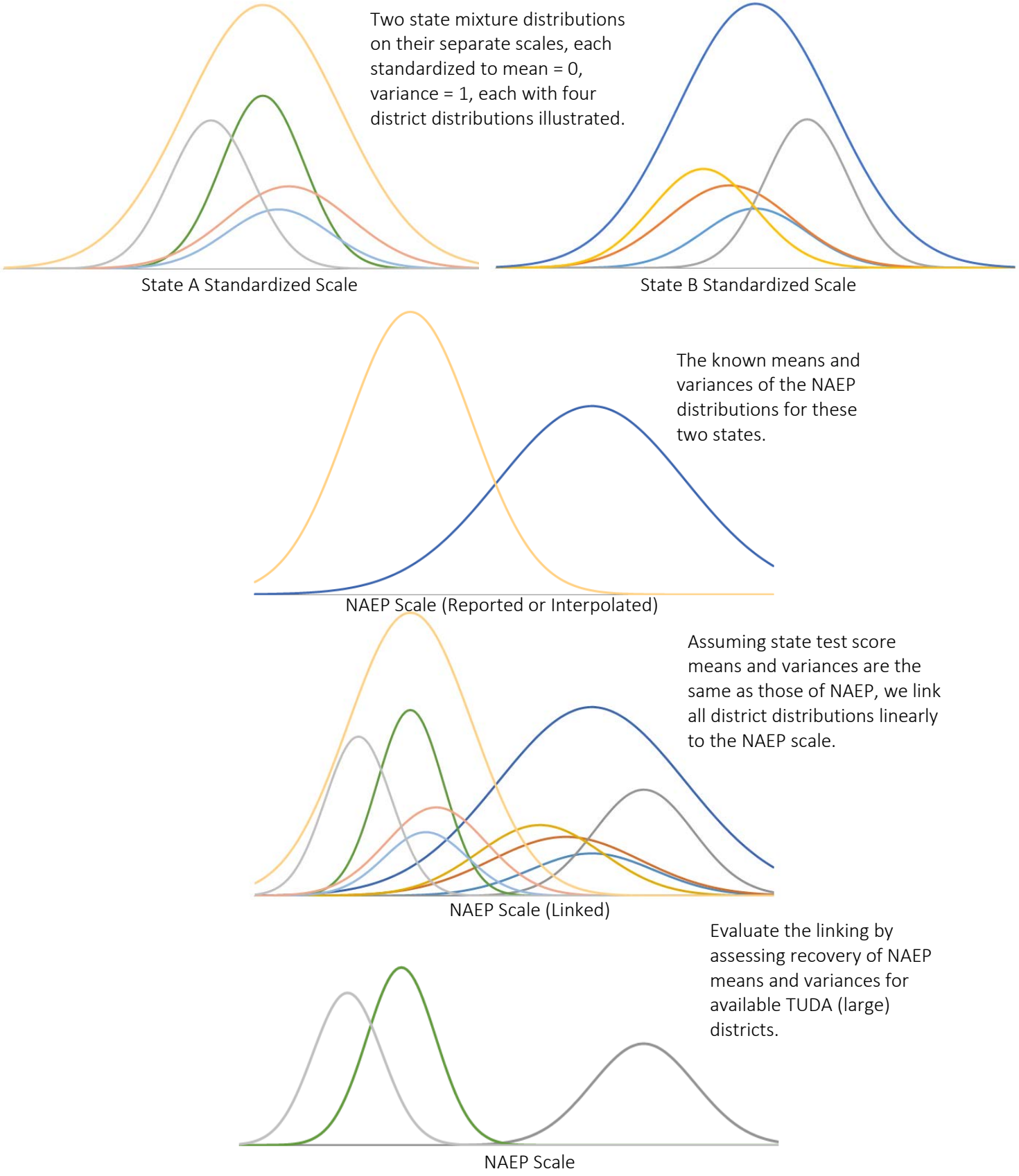


Figure 2: Comparing reported means from NAEP TUDA and NAEP-linked state test score distributions, grades 4 and 8, Reading and Mathematics, in 2009, 2011, and 2013. Districts and years with a greater than 8-point discrepancy are labeled.

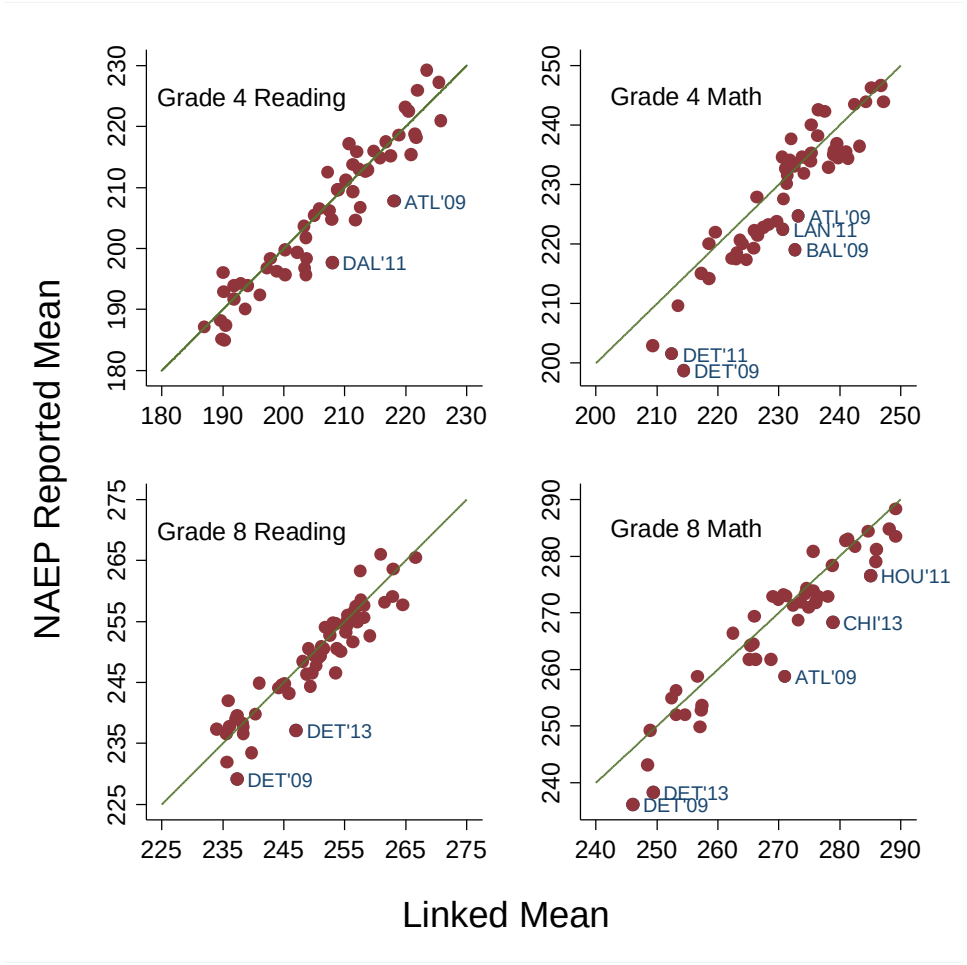
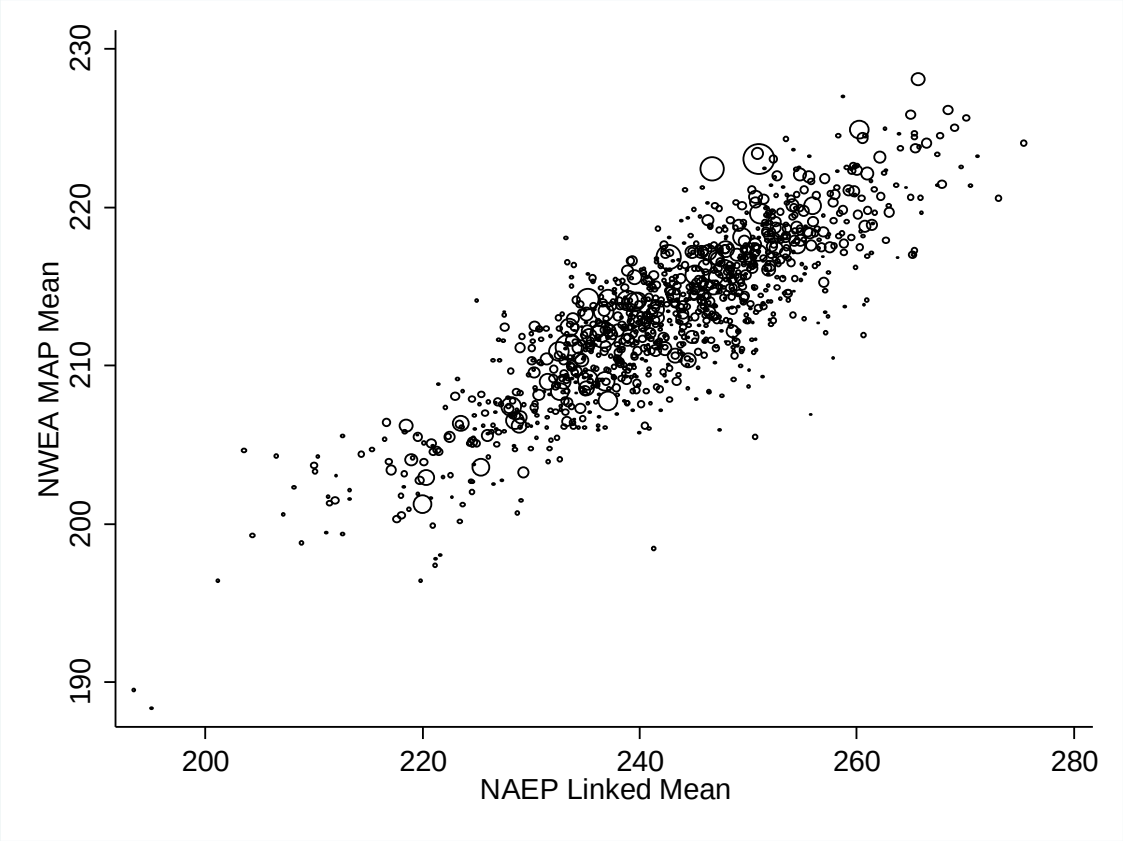


Figure 3. Example of an association between linked means and NWEA MAP means, Grade 4 Math, 2009.



Note: Correlation of .87; precision-adjusted correlation of .93. Bubble size corresponds to district enrollment.

Appendices

Table A1: Recovery of NAEP TUDA standard deviations following state-level linkage of state test score distributions to the NAEP scale.

			Recovery			Relationship	
Subject	Grade	Year	n	RMSE	Bias	Corr.	Adj. Corr.
Reading	4	2009	17	2.37	-1.23	0.53	0.76
		2011	20	2.62	-0.06	0.36	0.46
		2013	20	1.57	0.30	0.85	0.96
	8	2009	17	2.90	-1.63	0.23	0.89
		2011	20	2.68	-1.16	0.32	0.40
		2013	20	2.42	-1.33	0.50	0.62
Math	4	2009	17	1.60	0.06	0.58	0.72
		2011	20	1.90	0.89	0.6	0.68
		2013	20	2.03	1.18	0.48	0.71
	8	2009	14	2.62	-0.78	0.62	0.77
		2011	17	2.27	0.25	0.66	0.80
		2013	17	2.62	0.37	0.56	0.63
Average		2009	65	2.37	-0.89	0.49	0.79
		2011	77	2.37	-0.02	0.49	0.59
		2013	77	2.16	0.13	0.60	0.73
		Reading	114	2.43	-0.85	0.47	0.68
		Math	105	2.17	0.33	0.58	0.72
		All	219	2.30	-0.26	0.52	0.70

Note: Using NAEP Expanded Population Estimates. Adjusted correlations account for imprecision in linked and target estimates.

Table A2: Correlations with NWEA MAP district standard deviations before and after state-level linkage of state test score distributions to the NAEP scale.

				Correlations		Precision-adj. corr.	
Subject	Grade	Year	n	Pre-Link	Post-Link	Pre-Link	Post-Link
Reading	4	2009	1134	0.21	0.24	0.50	0.59
		2011	1476	0.29	0.40	0.56	0.65
		2013	1820	0.20	0.26	0.61	0.66
	8	2009	946	0.25	0.26	0.56	0.60
		2011	1276	0.27	0.35	0.54	0.60
		2013	1584	0.16	0.17	0.52	0.53
Math	4	2009	1123	0.26	0.34	0.67	0.77
		2011	1467	0.22	0.30	0.60	0.67
		2013	1831	0.20	0.23	0.66	0.72
	8	2009	959	0.21	0.24	0.66	0.74
		2011	1287	0.34	0.39	0.59	0.66
		2013	1544	0.24	0.29	0.67	0.73
Average		2009	4162	0.23	0.27	0.60	0.67
		2011	5506	0.28	0.36	0.57	0.64
		2013	6779	0.20	0.24	0.62	0.66
		All Years	16447	0.24	0.29	0.60	0.66

Note: Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Only includes districts with >90% of enrollment reporting scores on NWEA MAP.

Table A3. Recovery of reported 2011 NAEP TUDA standard deviations following state-level linkage of state test score distributions to a NAEP scale interpolated between 2009 and 2013.

			Recovery			Relationship	
Subject	Grade	Year	n	RMSE	Bias	Corr.	Adj. Corr.
Reading	4	2011	20	2.61	0.38	0.40	0.52
	8	2011	20	2.98	-0.80	0.02	-0.01
Math	4	2011	20	2.35	1.46	0.45	0.53
	8	2011	17	2.07	0.82	0.72	0.85
		Average	77	2.50	0.46	0.40	0.48

Note: Using NAEP Expanded Population Estimates. Adjusted correlations account for imprecision in linked and target estimates.

Table A4: Precision-adjusted correlations between NWEA MAP district standard deviations and NAEP-linked estimates.

Subject	Grade	2009	2010	2011	2012	2013
Reading	3	0.55	0.61	0.68	0.66	0.67
	4	0.59	0.60	0.65	0.64	0.66
	5	0.63	0.57	0.63	0.67	0.66
	6	0.62	0.61	0.63	0.67	0.63
	7	0.61	0.60	0.61	0.57	0.55
	8	0.60	0.57	0.60	0.55	0.53
Math	3	0.72	0.73	0.70	0.70	0.66
	4	0.77	0.80	0.67	0.72	0.72
	5	0.74	0.74	0.75	0.75	0.77
	6	0.77	0.78	0.70	0.74	0.74
	7	0.77	0.67	0.71	0.75	0.72
	8	0.74	0.67	0.66	0.71	0.73

No interpolation	0.659	Reading	0.61
Single interpolation	0.672	Math	0.73
Double interpolation	0.677		

Note. Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Only includes districts with >90% of enrollment reporting scores on NWEA MAP.

Table A5: Reliabilities of average NAEP state standard deviations over grades (4, 8) and years (2003, 2005, 2007, 2009, 2011, 2013)

	1 grade-year	2 grades	2 years	2 grades, 2 years
Reading	0.49	0.63	0.57	0.71
Math	0.70	0.80	0.78	0.86