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

AUTHORS

Sean F. Reardon
Stanford University

Demetra Kalogrides
Stanford University

Andrew D. Ho
Harvard University

ABSTRACT

There is no comprehensive database of U.S. district-level test scores that is comparable across states. We describe and evaluate a method for constructing such a database. First, we estimate linear, reliability-adjusted linking transformations from state test score scales to the scale of the National Assessment of Educational Progress (NAEP). We then develop and implement direct and indirect validation checks for linking assumptions. We conclude that the linking method is accurate enough to be used in analyses of national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions among districts in different states invalid. Finally, we describe several different methods of scaling and pooling the linked scores to support a range of secondary analyses and interpretations.

Acknowledgements: 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.

VERSION

June 2017

Suggested citation: Reardon, S.F., Kalogrides, D., & Ho, A. (2017). Linking U.S. School District Test Score Distributions to a Common Scale (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

Sean F. Reardon
Demetra Kalogrides
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

June, 2017

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 ED*Facts* 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

Abstract

There is no comprehensive database of U.S. district-level test scores that is comparable across states. We describe and evaluate a method for constructing such a database. First, we estimate linear, reliability-adjusted linking transformations from state test score scales to the scale of the National Assessment of Educational Progress (NAEP). We then develop and implement direct and indirect validation checks for linking assumptions. We conclude that the linking method is accurate enough to be used in analyses of national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions among districts in different states invalid. Finally, we describe several different methods of scaling and pooling the linked scores to support a range of secondary analyses and interpretations.

Keywords: *linking, scaling, multilevel modeling, achievement testing, NAEP*

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. We also describe methods of scaling and pooling the district-level test score estimates to provide interpretable and useful summary statistics for each district.

Existing and future assessments enable some district-level comparison across states, but none is comprehensive. Some data are available for comparisons of average test scores across states and districts. At the highest 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 reading/English Language Arts (ELA) tests federally mandated by the No Child Left Behind (NCLB) act, administered annually in grades 3-8. Comparing academic achievement across state lines requires either that districts administer a common test, or that the scores on the tests can be linked. However, state accountability tests generally 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 ongoing rollout of common assessments developed by multistate assessment consortia (such as the Partnership for Assessment of Readiness for College and Careers, PARCC, and the Smarter Balanced Assessment Consortium, SBAC) will certainly increase comparability across states, but only to the extent that states use these assessments. Customization of content standards by states may also discourage the reporting of results on a common scale across states (PARCC allows for customization of 15% of content standards). 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 lower-stakes 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. However, the districts using MAP are neither a representative nor comprehensive 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) estimated state-level test score distributions from NAEP and population-level 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) several approaches to scaling and pooling the results for secondary interpretations and uses. None of the components of our approach is novel on its own, but together they represent a suite of approaches for developing, evaluating, scaling, and pooling linked estimates of test score distributions.

We use data from the ED*Facts* Initiative (U.S. Department of Education, 2015), NAEP, and NWEA. We estimate the necessary transformations using a) an application of heteroskedastic 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. Once we link the means and standard deviations to the NAEP scale, we show how they can be standardized to facilitate comparisons of districts within grade-year cells, across years, and across grades. We also describe a model for pooling estimates across grades, years, and subjects within a given district to simplify between-district comparisons.

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 & McGee, 2011) and school level (Greene & Mills, 2014), the *Global Report Card* maps scores onto a national scale using proficiency rates, using a somewhat different approach than ours.[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 we do not intend 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, using a variety of validation checks to

[1] Our data and methods are more comprehensive than the *Global Report Card* (Greene & McGee, 2011; Greene & Mills, 2014; http://globalreportcard.org/). First, we provide grade-specific estimates (by year), allowing for estimates of measures of progress. Second, 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. Third, our methods allow us to provide standard errors for our estimates. Fourth, we provide both direct and indirect validation checks for our linkages.

assess the extent to which our methods yield unbiased estimates of aggregate means and standard deviations.

Data

We use state accountability test score data and NAEP data to link scores, and we use NAEP data and NWEA MAP data to evaluate the linkage. Under the ED*Facts* Initiative (U.S. Department of Education, 2015), states report frequencies of students scoring in each of several 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, most commonly 4. We use ED*Facts* 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 are not suppressed and have no 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 deviations 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 probability samples of students in the relevant grades and years. We use years 2009, 2011, and 2013 as a basis for linking; we use additional

[2] We add random error $e_i \sim N(0, \omega_i^2/n_i)$ to each estimate reported, where ω_i^2 is the sampling variance of the parameter estimate (e.g., a district's estimated mean or standard deviation) in district-subject-grade-year cell i, and n_i is the number of students tested in that 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.

odd years from 2003 through 2007 as part of some validation analyses. The NAEP state and 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, & 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 ED*Facts* data: 2009-2013, grades 3-8, in reading/ELA 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 ED*Facts* 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. These comprise roughly 10% of the districts using the NWEA MAP tests. After these exclusions, we estimate district-grade-subject-year 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 ED*Facts* data to district means and standard deviations expressed

[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.

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, 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., ED*Facts*). 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 ED*Facts* 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 $\hat{\mu}_{dygb}^{\text{state}}$ and $\hat{\sigma}_{dygb}^{\text{state}}$, respectively, for district d, year y, grade g, and subject b. The HETOP model estimation procedure also provides standard errors of these estimates, denoted $se\left(\hat{\mu}_{dygb}^{\text{state}}\right)$ and $se\left(\hat{\sigma}_{dygb}^{\text{state}}\right)$, 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 $\hat{\mu}_{dygb}^{\text{state}}$) to its corresponding NAEP distribution. Recall that we have estimates of NAEP means and standard deviations at the state (denoted s) level, denoted $\hat{\mu}_{sygb}^{\text{naep}}$ and $\hat{\sigma}_{sygb}^{\text{naep}}$, 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 between grades 4 and 8 to grades 5, 6, and 7 and extrapolate to grade 3.

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.

$$\hat{\mu}^{naep}_{sygb} = \hat{\mu}^{naep}_{sy4b} + \frac{g-4}{4}\left(\hat{\mu}^{naep}_{sy8b} - \hat{\mu}^{naep}_{sy4b}\right), \quad \text{for g} \in \{3,5,6,7\}$$
$$\hat{\mu}^{naep}_{sygb} = \frac{1}{2}\left(\hat{\mu}^{naep}_{s[y-1]gb} + \hat{\mu}^{naep}_{s[y+1]gb}\right), \quad \text{for y} \in \{2010,2012\} \tag{1}$$

We evaluate the validity of linking to interpolated NAEP grades and years explicitly later 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 $\hat{\mu}^{\widehat{\text{naep}}}_{dygb}$. 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).

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

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

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 and standard deviations on the state scale, $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$, are expressed in terms of standard deviation units of the state score distribution. The standardized means are attenuated toward zero due to measurement error. They must be disattenuated before being mapped to the NAEP scale, given that the NAEP scale accounts for measurement error due to item sampling. We disattenuate the means by dividing them by the square root of the state test score reliability estimate, $\hat{\rho}^{\text{state}}_{sygb}$. The district standard deviations on the state scale, $\hat{\sigma}^{state}_{dygb}$, 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 for non-interpolated grades and years:

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

for g ∈ {4,8} and y ∈ {2009,2011,2013};

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

for g ∈ {4,8} and y ∈ {2009,2011,2013}.

For interpolated grades and years, the sampling variances differ, because interpolated and extrapolated values are essentially weighted averages. For example, it follows from Equation (1), and

assuming uncorrelated terms, that for grades 3, 5, 6, and 7 in odd years the sampling variances of interpolated and extrapolated means are:

$$var\left(\hat{\mu}_{dygb}^{\widehat{naep}}\right) = \left(\frac{8-g}{4}\right)^2 var\left(\hat{\mu}_{dy4b}^{\widehat{naep}}\right) + \left(\frac{g-4}{4}\right)^2 var\left(\hat{\mu}_{dy8b}^{\widehat{naep}}\right), \text{for g} \in \{3,5,6,7\} \quad (6)$$

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 districts participating in the NAEP TUDA assessments over these years, we compare $\hat{\mu}_{dygb}^{\widehat{\text{naep}}}$—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 that the linking method attempts to recover but are not themselves used in any way in the linking process. We repeat this linkage for demographic subgroups to assess the population invariance of the link. Second, we assess the correlation of our linked district estimates with district mean scores on the NWEA MAP tests. This provides the correlation across a larger sample of districts. However, the NWEA MAP test has a different score scale, so it does not provide direct comparability with the NAEP scale that is the target of our linking. Third, for the 20 relevant TUDA districts, we assess whether within-district differences in linked scores across grades and cohorts correspond to those differences observed in the NAEP 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. 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 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.[5] 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 discrepancies $\widehat{\hat{\mu}_{dygb}^{\text{naep}}} - \hat{\mu}_{dygb}^{\text{naep}}$ and $\widehat{\hat{\sigma}_{dygb}^{\text{naep}}} - \hat{\sigma}_{dygb}^{\text{naep}}$. If there were no sampling or measurement error in these estimates, we would report the average of these discrepancies as the bias, and would report the square root of the average squared discrepancies as the Root Mean Squared Error (RMSE). We could also report the observed correlation between the two as a measure of how well the linked estimates align linearly with their reported TUDA values. However, 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. Instead, we report measurement-error corrected RMSEs and correlations that account for imprecision in both the linked and TUDA parameter estimates. To estimate the measurement-error corrected bias, RMSE, and correlation in a given year, grade, and subject, we fit the model below, using the sample of districts for which we have both estimates $\widehat{\hat{\mu}_{dygb}^{\text{naep}}}$

[5] 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 (and Texas districts in 2013) did not have a common Grade 8 state mathematics assessment, so the California and Texas districts lack a linked district mean for Grade 8 mathematics.

and $\hat{\mu}^{\text{naep}}_{dygb}$ (or $\widehat{\hat{\sigma}^{\text{naep}}_{dygb}}$ and $\hat{\sigma}^{\text{naep}}_{dygb}$ as the case may be; the model is the same for the means or standard deviations):

$$\hat{\mu}_{idygb} = \alpha_{0dygb}(LINKED_i) + \alpha_{1dygb}(TUDA_i) + e_{idygb}$$

$$\alpha_{0dygb} = \beta_{00} + u_{0dygb}$$

$$\alpha_{1dygb} = \beta_{10} + u_{1dygb}$$

$$e_{idygb} \sim N\left(0, \omega^2_{idygb}\right);\ \boldsymbol{u}_{dygb} \sim MVN(0, \boldsymbol{\tau}^2), \tag{7}$$

where i indexes source (linked or NAEP TUDA test), ω^2_{idygb} is the estimated sampling variance of $\hat{\mu}_{idygb}$, and $\boldsymbol{\tau}^2 = \begin{bmatrix} \tau^2_{00} & \tau^2_{01} \\ \tau^2_{01} & \tau^2_{11} \end{bmatrix}$ is the variance-covariance matrix of the linked and TUDA parameter values which must be estimated. Given the model estimates, we estimate the bias, $\hat{B} = \hat{\beta}_{00} - \hat{\beta}_{10}$, and $\widehat{RMSE} = \left[\hat{B}^2 + \mathbf{b}\hat{\boldsymbol{\sigma}}^2\mathbf{b}'\right]^{1/2}$ where $\mathbf{b} = [\mathbf{1} - \mathbf{1}]$ is a 1×2 design matrix. Finally, we estimate the correlation of α_{0dygb} and α_{1dygb} as $\hat{r} = \frac{\hat{\tau}^2_{01}}{\hat{\tau}_{00}\hat{\tau}_{11}}$.

Table 1 reports the results of these analyses in each subject, grade, and year in which we have TUDA estimates (see Table A1 for the corresponding table for standard deviations). Although we do not show the uncorrected estimates here, we note that the measurement error corrections have a negligible impact on bias and reduce the (inflated) RMSE by around 8% on average. On average, the linked estimates overestimate actual NAEP TUDA means by roughly 1.8 points on the NAEP scale, or around 0.05 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 1 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 1 also shows that the average estimated precision-adjusted correlation (disattenuated to account for the imprecision in the observed means) is 0.95 (note that the simple unadjusted correlation is 0.94; measurement error in the means is relatively minor relative to the true variation in the means of the TUDA districts). 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. If we had all districts, the bias in Table 1 would be 0; it is not 0 because Table 1 summarizes the bias for only the subset of NAEP urban districts for which we have scores. The RMSE similarly describes the magnitude of error (the square root of average squared error) for these districts and may be larger or smaller than the RMSE for other districts in the state.

We review here four 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 1 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 1 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 1 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 (Wilson, Bowers, & Hyde, 2011). 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.

We also address the population invariance of the linking (e.g., Kolen & Brennan, 2014; Dorans & Holland, 2000) by reporting the average direction and magnitude (RMSE) of discrepancies, $\widehat{\hat{\mu}^{\mathrm{naep}}_{dygb}} - \hat{\mu}^{\mathrm{naep}}_{dygb}$, for selected gender and racial/ethnic subgroups in Table 1. The numbers of districts is lower in

some grade-year cells due insufficient subgroup samples in some districts.[6] The RMSEs are only slightly larger for subgroups than the RMSE for all students, and bias is around the same magnitude, albeit smaller for White students than for the other subgroups. We conclude from these comparable values that the linking functions recover NAEP district means similarly, on average, across subgroups.

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 "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. The "pre-link" correlation is one where states differ on the MAP scale but do not, on average, on the state scale. The improvement from this correlation to the "post-link" correlation is due solely to the move from the "state" to the "$\widehat{\mathrm{naep}}$" scale, shifting all districts within each state according to NAEP performance. For means:

[6] Our model for subgroups pools across grades (4 and 8) and years (2009, 2011, and 2013) to compensate for smaller numbers of districts in some grade-year cells. We describe this model in Validation Check 3 below. On average across grades and years, the results are similar to a model that does not use pooling. We also calculate bias and RMSE for Asian student populations but do not report them due to small numbers of districts: 5-10 per cell. However, bias and RMSE of linked district estimates were higher for Asians, suggesting caution against conducting a separate linkage for Asian students.

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

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

(8)

Table 2 shows that correlations between "post-link" district means and MAP district means are 0.93 when adjusting for imprecision (see Table A2 for the corresponding table for standard deviations). These "post-link" correlations are larger than the "pre-link" correlations of 0.86. Figure 3 shows a bubble plot of district MAP scores on linked scores for Grade 4 mathematics in 2009, as an illustration of the data underlying these correlations. Note that the points plotted in Figure 3 are means estimated with imprecision. The observed (attenuated) correlations are generally .03 to .10 points lower than their disattenuated counterparts.

Table 2 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 district means compares not just the means in a given grade and year, but 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 years and grades, even when there is a small bias in in the average means.

To assess the accuracy of the across-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 the same model for both means and standard deviations, and separately by subject. For each model, we fit precision-weighted random coefficients models of this form:

$$\hat{\mu}_{idygb} = \alpha_{0dygb}(LINKED_i) + \alpha_{1dygb}(TUDA_i) + e_{idygb}$$

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

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

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

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

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

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

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

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

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

(9)

where i indexes source (linked or NAEP TUDA test) and ω^2_{idygb} is the sampling variance of $\hat{\mu}_{idydb}$ (which we set equal to the square of its estimated standard error). The vector $\boldsymbol{\Gamma} = \{\gamma_{00}, \ldots, \gamma_{12}\}$ contains the average intercepts, year slopes, and grade slopes (in the second subscript, 0, 1, and 2, respectively) for the linked values and the target values (in the first subscript, 0 and 1, respectively). The differences between the corresponding elements of $\boldsymbol{\Gamma}$ 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 1 above, where we estimated bias separately for each year and grade and descriptively averaged them, the bias here is estimated by pooling over all years and grades of TUDA data, with district random effects. If the linking were perfect, we would expect this to be 0.

The matrix of random parameters $\boldsymbol{\tau}^2$ 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 year-to-year 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 linked and TUDA mean scores. This correlation differs from that shown in Table 1 above because the former estimates the correlation separately for each grade and year; the model in Equation 9 estimates the correlation from a model in which all years and grades are pooled.

Table 3 shows the results of fitting this model separately by subject to the district means (see Table A3 for the corresponding table for standard deviations). 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 3) is small and not statistically significant. The linked reading 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 mathematics means are, on average, 2.4 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 1.

Table 3 here

Second, the estimated average linked and TUDA grade slopes ($\hat{\gamma}_{02}$ and $\hat{\gamma}_{12}$, respectively) are nearly identical to one another; this is true in both math and reading. The estimated bias in grade slopes (-0.04 in reading and -0.10 in math) is only 1% as large as the average grade slope. The implied RMSE from the model is 0.56 in reading and 0.46 in math, roughly 5% of the average grade slope.[7] The estimated correlation of the TUDA and linked grade slopes is 0.85 for reading and 0.98 for math. Finally,

[7] We compute the RMSE of the grade slope from the model estimates as follows. Let $C = \gamma_{02} - \gamma_{12}$ be the bias in the grade slopes; then the RMSE of the grade slope will be: $RMSE = [C^2 + \mathbf{d}\boldsymbol{\tau}^2\mathbf{d}']^{1/2}$, where $\mathbf{d} = [0\ 0\ 1\ 0\ 0 - 1]$.

the reliability of the grade slopes of the linked estimates is 0.76 in reading and 0.73 in math.[8] Together these indicate that the linked estimates provide unbiased estimates of the within-district differences 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 years that are sufficiently reliable to be useful (in models not shown, we estimate the reliabilities of the TUDA year trends to be 0.28 and 0.53 and of the linked year trends to be 0.45 and 0.72 in reading and math, respectively). As a result, we dropped the random effects on the year trends and do not report in Table 3 estimates of the variance, correlation, or reliability of the year trends.

Validation Check 4: Recovery of estimates under interpolation within years

Using the interpolated state means and standard deviations in Equation (1) for the linking establishes an assumption that the linkage recovers district scores that would have been reported in years 2010 and 2012 and grades 3, 5, 6, and 7. 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 state 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] The reliability of the level-3 slopes and intercepts is computed as described in Raudenbush and Bryk (2002).

(10)

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 bias and RMSE for discrepancies $\hat{\mu}^{\widehat{\text{naep}}'}_{d2011gb} - \hat{\mu}^{\text{naep}}_{d2011gb}$ using the model from Validation Check 1. Table 4 shows results in the same format as Table 1 (see Table A4 for the corresponding table for standard deviations). We note that the average RMSE of 3.8 and bias of 1.4 in Table 4 are approximately the same as the average RMSE of 3.8 and bias of 1.6 shown for 2011 in Table 1. Note that the interpolations we use in our proposed linking are those between observed scores that are only two years apart, rather than four years apart as in the validation exercise here. The two-year interpolations should be 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 4 here

We next investigate the viability of interpolation by comparing correlations of linked district estimates with MAP scores at different degrees of interpolation. Some grade-year combinations need no interpolation, others are singly interpolated, and others are doubly interpolated. Table 5 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 (see Table A5 for the corresponding table for standard deviations). This lends additional evidence that interpolation adds negligible aggregate error to recovery.

Table 5 here

Scaling

The linked estimates, $\hat{\mu}^{\widehat{\text{naep}}}_{dygb}$ and $\hat{\sigma}^{\widehat{\text{naep}}}_{dygb}$, of district test scores are expressed on the NAEP math and reading scales, which are comparable over time. Here, we present four scaling techniques to linearly

transform estimates from the NAEP scale to metrics that may be more suitable for analysis and interpretation. These techniques each allow a different scope of comparison across years and grades. We describe them here, with notation for district means in the following bullets for illustration:

- $\hat{\mu}^{\widehat{n*}}_{dygb}$ (national cell), a scale referenced to grade-, year-, and subject-specific national student-level standard deviation units. Absolute comparisons are possible across states but not grades, years, and subjects.
- $\hat{\mu}^{\widehat{y*}}_{dygb}$ (year), a scale referenced to grade- and subject-specific national student-level standard deviation units for a given year, in our case: 2009. Absolute comparisons are possible across states and years but not grades and subjects.
- $\hat{\mu}^{\widehat{c*}}_{dygb}$ (cohort), a scale referenced to grade- and subject-specific national student-level standard deviation units for a given cohort, in our case, the 2009 grade 4 cohort (that was in grade 5 in 2010, etc.). Like $\hat{\mu}^{\widehat{y*}}_{dygb}$, absolute comparisons are possible across states and years but not grades and subjects.
- $\hat{\mu}^{\widehat{g*}}_{dygb}$ (grade), a scale referenced to units of grade-level differences in average scores, in our case, for the 2009 grade 4 cohort: the difference between their 2009 grade 4 and 2013 grade 8 average scores. Absolute comparisons are possible across states, years, and grades but not subjects.

First, we apply Equation 1 to national NAEP scores, resulting in a table of interpolated estimates of the national means and standard deviations of achievement on the NAEP scale for each grade and year. We denote these $\hat{\mu}^{\mathbf{naep}}_{ygb}$ and $\hat{\sigma}^{\mathbf{naep}}_{ygb}$, respectively. These national means and SDs differ from their district- and state-level estimates by lacking the subscript, d or s. The results of this process are shown in Table 6.

Table 6 here

We can standardize the linked estimates to their respective cells in Table 6: the national student-level distribution of NAEP scores in each year, grade, and subject. That is, we compute:

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

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

(11)

The problem with this method of standardization is that it destroys information about real changes over time and real differences across grades, inferences that are enabled by the stability and vertical linkages of the NAEP scale.

An alternative is to standardize the scores using a particular mean and standard deviation from a single year, denoted $[y^*]$. This rightfully allows for absolute comparisons across years enabled by the stable NAEP scale. In Table 6, we highlight in light gray the year, $y^* = 2009$, to illustrate the national means and standard deviations to which we could standardize all district means, across years, within a given grade:

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

$$\hat{\sigma}^{\widehat{y*}}_{dygb} = \frac{\hat{\sigma}^{\widehat{\text{naep}}}_{dygb}}{\hat{\sigma}^{\text{naep}}_{[y^*]gb}}$$

(12)

A third alternative is to standardize the scores using the means and standard deviations of a particular cohort, denoted $[c^*]$. We highlight these cells in dark gray in Table 6 for $c^* = 2005$, the cohort in kindergarten in 2005 that is also in Grade 4 in 2009 and Grade 8 in 2013.

$$\hat{\mu}^{\widehat{c*}}_{dygb} = \frac{\hat{\mu}^{\widehat{\text{naep}}}_{dygb} - \hat{\mu}^{\text{naep}}_{[(y,g)^*]b}}{\hat{\sigma}^{\text{naep}}_{[(y,g)^*]b}}, for\ (y,g)^*\ s.t. y - g = c^*$$

$$\hat{\sigma}_{dygb}^{\widehat{c^*}} = \frac{\hat{\sigma}_{dygb}^{\widehat{\text{naep}}}}{\hat{\sigma}_{[(y,g)^*]b}^{\text{naep}}}, for\ (y,g)^*\ s.t. y - g = c^*$$

(13)

The district test score distributions are now standardized using the estimated grade-specific national student-level distribution of scores from a common cohort of students. Other grade-specific estimates of means and standard deviations may be reasonable, including those from an earlier year, and earlier cohort, or averages across multiple years and cohorts. Table 6 shows that these magnitudes are fairly similar across choices of c^* and y^*. Conversions among different standardizations are straightforward as long as the choice of cell or averages among cells is clear.

A fourth way of standardizing the estimates is to convert them to approximate units of average between-grade score differences. To do this, we first estimate the within-cohort change in subject b, for the same cohort of students in kindergarten in 2005, $c^* = 2005$, by using estimates of the national NAEP means and standard deviations in grade 8 in 2013 and grade 4 in 2009. This is denoted $\hat{\gamma}_{c^*b}$,[9] e.g.,:

$$\hat{\gamma}_{2005b} = \frac{\hat{\mu}_{2013,8b}^{\text{naep}} - \hat{\mu}_{2009,4b}^{\text{naep}}}{4}$$

(14)

We then identify the linear transformation that sets these grade 4 and 8 averages at the "grade level" values 4 and 8 respectively, and transform all other district scores accordingly:

$$\hat{\mu}_{dygb}^{\widehat{g^*}} = 4 + \frac{\hat{\mu}_{dygb}^{\widehat{\text{naep}}} - \hat{\mu}_{2009,4b}^{naep}}{\hat{\gamma}_{c^*b}},$$

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

(15)

[9] It is also possible to define a grade level as $\hat{\gamma}_{y^*b}$, a quarter of the difference between grade 4 and 8 scores within a given year, instead of within a given cohort. As calculated from Table 6, the difference is negligible, transparent, and readily transferable to other cross-grade reference points, including other cohorts, years, and averages among these.

On this basis, $\hat{\mu}^{\widehat{g*}}_{dygb}$ can be interpreted as the estimated average national "grade-level performance" of students in district d, year y, grade g, and subject b. So if $\hat{\mu}^{\widehat{g*}}_{dy4b} = 5$, students in district d, year y are one grade level ($\hat{\gamma}_{c^*b}$) above the 4th grade 2009 national average ($\hat{\mu}^{naep}_{2009,4b}$) in performance on the tested subject b.

The four methods of standardization enable different interpretations. The first (denoted by the superscript $\hat{n}^*$) 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 $\hat{y}^*$) expresses districts' score distributions in units of the grade-specific national standard deviation units in a specific year. The scale compares a district's average achievement in a given grade and year to the national average in that grade in some reference year. This scale retains information about absolute changes over time, and 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 seems appropriate, as our analyses above show. This scale describes relative differences in districts' scores within grades, but does not provide information about absolute differences across grades, because the scale is standardized within each grade.

The third method of standardization (denoted $\hat{c}^*$) expresses districts' score distributions in units of a given cohort's grade-specific national standard deviation units. The scale compares a district's average achievement in a given grade and year to the national average in that grade in the year when a specific cohort was in that grade. Like the $\hat{y}^*$ scale, this scale retains information about absolute changes

over time by relying on the stability of the NAEP scale over time and on the linear interpolation of NAEP distributions over time. Likewise, this scale does not enable absolute comparisons across grades. The key distinction between the $\hat{c}^*$ and the $\hat{y}^*$ scales is that the first is standardized to a specific year and the second to a specific cohort. Depending on the use of the scale, one may be more preferable than the other, though they will in general be very similar, as calculations from Table 6 can show.

Finally, the fourth method of standardization (denoted $\hat{g}^*$) expresses district score distributions in units that correspond to national grade-level averages. On this scale, a one-unit difference corresponds to a national average within-cohort difference in scores between students in adjacent grades. The scale is set so that a value of 4 corresponds to the average NAEP score among 4th graders in the cohort of students who were in 4th grade in 2009; a value of 8 corresponds to the average NAEP score among 8th graders in 2013. Other anchors are possible and make modest differences identifiable in Table 6. This metric enable absolute comparisons across grades and over time, but it does so by relying on the linear interpolation of NAEP scores between grades and years, on the assumption that the NAEP scale is stable over time, and on the assumption that the NAEP scale is vertically linked across grades 4 and 8.

Vertical linking was built into NAEP's early design via cross-grade blocks of items (administered to both 4th and 8th graders) in 1990 in mathematics and in 1992 in reading (Thissen, 2012). These cross-grade blocks act as the foundation for the bridge spanning grades 4 and 8. At around that time, the National Assessment Governing Board that sets policy for NAEP adopted the position that, as Haertel (1991) describes, "NAEP should employ within-age scaling whenever feasible" (p. 2). Thissen notes that there have therefore been few checks on the validity of the cross-grade scales since that time. One exception is a presentation by McClellan, Donoghue, Gladkova, and Xu (2005), who tested whether subsequent vertical linking would have made a difference on the reading assessment. They concluded that "the current cross-grade scale design used in NAEP seems stable to the alternate design studied" (p. 37). We include the grade scale because it amounts essentially to a choice of a linear scaling constant and

is more readily interpretable, particularly to non-technical audiences. However, we do not advise it for analyses where the vertical linking across grades and the linear interpolation assumptions are not required or defensible.

Pooling estimates within districts

The procedures described above yield estimates of the means and standard deviations of school districts' test score distributions on a common scale across states, grades, and years. In some cases, it will be useful to pool the estimates within each district to provide summary measures of overall academic performance and average changes in performance across grades and years. For example, we may want to characterize a school district not only by its students' average level of performance, but also by the linear trend in their performance across cohorts of students (within the same grade) or across grades (within a given cohort).

For each district, we have up to 30 grade-and year-specific estimated means and standard deviations per subject (spanning 6 grades and 5 years). We define a cohort variable ($cohort = year - grade$) that is equal to the year that a cohort would have been in the spring of their kindergarten year (so our oldest cohort of students—those in 8th grade in spring 2009--are in cohort 2001; our youngest cohort—3rd-graders in 2013—are in the 2010 cohort). In the previous section, we describe procedures for standardizing these in various ways. For any given standardization metric (denoted with the superscript $\hat{x}^*$, which may signify the $\hat{n}^*$, $\hat{y}^*$, $\hat{c}^*$, or $\hat{g}^*$ metric) and subject b, we can fit the precision-weighted random coefficients model:

$$\hat{\mu}_{dygb}^{\hat{x}^*} = \beta_{0d} + \beta_{1d}\left(cohort_{dygb} - 2005.5\right) + \beta_{2d}\left(grade_{dygb} - 5.5\right) + u_{dygb} + e_{dygb}$$

$$\beta_{0d} = \gamma_{00} + v_{0d}$$

$$\beta_{1d} = \gamma_{10} + v_{1d}$$

$$\beta_{2d} = \gamma_{20} + v_{2d}$$

$$e_{dygb} \sim N\left(0, \omega^2_{dygb}\right);\ u_{dygb} \sim N(0, \sigma^2);\ \begin{bmatrix} v_{0d} \\ v_{1d} \\ v_{2d} \end{bmatrix} \sim MVN(0, \boldsymbol{\tau}^2),$$

(16)

In this model, β_{0d} represents the mean test score in subject b, in district d, in grade 5.5 for kindergarten cohort 2005.5 (this pseudo-cohort and psuedo-grade represents the center of our data's grade and cohort ranges). Note this differs from our choices in Equations 9 and 13, as our purpose here is to center the intercept between linked grades 3-8 instead of NAEP-observed grades 4 and 8. The β_{1d} parameter indicates the average within-grade (cohort-to-cohort) change per year in average test scores in district d; and β_{2d} indicates the average within-cohort change per grade in average test scores in district d. If the model is fit using one of the scales that standardizes scores within grades (the $\hat{n}^*$, $\hat{y}^*$,or $\hat{c}^*$ scales), the coefficients will be interpretable in NAEP student-level standard deviation units (relative to the specific standard deviation used to standardize the scale). Between-district differences in β_{0d}, β_{1d}, and β_{2d} will be interpretable relative to this same scale. If the model is fit using the grade-level scale ($\hat{g}^*$), the coefficients will be interpretable as test score differences relative to the average between-grade difference among students. As a result, for example, the average district-level mean (γ_{00}) will be near 5.5 and the average change per grade (γ_{20}) will be near 1, since the scale is standardized so that, on average, scores differ by 1 unit per grade.

Table 7 reports the results of fitting the model in Equation (16) to the district-level means data (See Table A6 in the Appendix for the corresponding results for standard deviations). We show results using the data standardized in the $\hat{c}^*$ and $\hat{g}^*$ scales (the results are very similar in the other scales). As expected, the average district intercepts are roughly 0 and 5.5 in the $\hat{c}^*$ and $\hat{g}^*$ scales, respectively.[10] In the $\hat{c}^*$ scale, test scores increase, in the average district, by 0.017 standard deviations per year. In the

[10] They are not exactly 0 and 5.5 for two reasons. First, the scales are standardized using NAEP data from a specific cohort, whereas the model is fit to all cohorts. Second, district estimates are not weighted by district size in the regression model, so the intercept acts as an average of averages. In contrast, the standardizations are referenced to national distributions, as if they were based on weighted means.

$\hat{g}^*$ scale, average scores increase by roughly 0.05 grade levels from cohort to cohort. This is consistent with the national gains seen in Table 6 when converted to grade equivalents using Equation 15. Because both the $\hat{c}^*$ and $\hat{g}^*$ scales are standardized around national grade-specific means; the average grade slopes are not particularly informative: they indicate that test scores change by an average of roughly 0 or 1 units per grade within a cohort, in the $\hat{c}^*$ and $\hat{g}^*$ scales, respectively, as expected.

Table 7 here

The other useful parameters in Table 7 are the estimated variance components and reliabilities. First, note that the between-district variance in test score means is very large relative to the residual within-district variance: 91-95% of the variance in means is between districts (similarly, in Table A6, 87-89% of the variance in standard deviations is between districts). This indicates the district-specific intercepts and linear grade and cohort trends describe almost all of the variation in test score distributions in our data. Moreover, the variances of the grade and cohort slopes are statistically significant and indicate substantial variation among districts. For example, the math grade slope has a variance of 0.004 in the $\hat{c}^*$ scale. That implies that the difference in average math test score growth from grade 3 to 8 between the districts with the fastest and slowest growth in scores is roughly 1.20 standard deviations.[11] The variation in reading score growth is roughly two-thirds as large as in math. Variation of these magnitudes is quite substantial. Similarly, Table A6 shows that the variation in the grade slopes on test score standard deviations is also quite substantial.

Finally, Table 7 indicates that the model provides very reliable estimates of differences among school districts in test score averages (reliabilities of 0.97). The district-specific grade and cohort slopes are slightly less reliably estimated (reliabilities are 0.65 to 0.75 for means), but are still reliable enough to

[11] The math grade slope variances implies a standard deviation of grade slopes, $\hat{\tau}_2$, of 0.061 for the $\hat{c}^*$ scale. A school district with math test score growth rates 2 standard deviations above the mean would have a growth rate of roughly +0.12 standard deviations per grade, or +0.60 standard deviations between grades 3 and 8. A school district with math test score growth rates 2 standard deviations below the mean would have a growth rate of roughly -0.12 standard deviations per grade or -0.60 standard deviations between grades 3 and 8.

meaningfully distinguish large and medium-sized districts from one another. In sum, Table 7 indicates that pooling the estimates across grades and cohorts can be a useful way to summarize each district's test score distributions and the variation of these distributions across grades and cohorts. Indeed, Table 7 suggests that the 30 estimates of a district's mean test scores can be summarized in three parameters: the average scores in a district, the linear trend in these scores across cohorts, and the linear trend in average scores across grades within a cohort. In some cases, one might wish to use the empirical Bayes shrunken estimates of these parameters, which can be computed after fitting Equation (16).

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 appropriate 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). Nonetheless, we find some evidence that NAEP-linked estimates include some small, but systematically positive, bias in large urban districts (roughly +0.05 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. Of course, within-state

comparisons do not depend on the linking procedure, so these are immune to bias and variance that arises from the linking methods.

The linked estimates are, we believe, accurate enough to be used to investigate broad patterns in the relationships 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 within-grade trends over time, given that there is little or no variation in the NAEP districts' trends over time against which to benchmark the linked trend estimates. This is true more generally even of within-grade national NAEP trends, which are often underpowered to detect true progress over shorter time spans of 2- to 4-years.

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, had we observed them, would differ. But our validation checks suggest that they would be much 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 set of estimated NAEP-linked district test score means, which is unprecedented in its scope and geographical detail, may be 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. (2010). An investigation of bias in reports of the National Assessment of Educational Progress. *Educational Evaluation and Policy Analysis, 32,* 24-43.

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.

Greene, J. P., & McGee, J. M. (2011). George W. Bush Institute Global Report Card: Technical Appendix. Report prepared for the George W. Bush Institute. Retrieved from http://www.globalreportcard.org/docs/AboutTheIndex/Global-Report-Card-Technical-Appendix-9-28-11.pdf

Greene, J. P., & Mills, J. N. (2014). George W. Bush Institute Global Report Card 3.0: Technical Appendix. Report prepared for the George W. Bush Institute. Retrieved from http://www.bushcenter.org/stateofourcities/data/GWBI_GRC_TechnicalAppendix.pdf

Haertel, E. H. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. Washington, DC: National Center for Educational Statistics.

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

McClellan, C. A., Donoghue, J. R., Gladkova, L., & Xu, X. (2005, November). *Cross-grade scales in NAEP: Research and real-life experience*. Presentation at the conference, Longitudinal Modeling of Student Achievement, Maryland Assessment Research Center for Education Success, University of Maryland, College Park, MD.

National Center for Education Statistics (2015). *Number of public elementary and secondary education agencies, by type of agency and state or jurisdiction: 2012-13 and 2013-14*. Table 214.30. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_214.30.asp?current=yes

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

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.

Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results.* NAEP Validity Studies Panel. Washington, DC: National Center for Education Statistics.

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

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.

Wilson, R. E., Bowers, M. J., & Hyde, R. L. (2011). *Special Investigation into Test Tampering in Atlanta's School System*. Retrieved from https://ia801000.us.archive.org/32/items/215252-special-investigation-into-test-tampering-in/215252-special-investigation-into-test-tampering-in.pdf

Table 1: Recovery of NAEP TUDA means following state-level linkage of state test score distributions to the NAEP scale, measurement error adjusted.

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2009	17	3.95	2.12	0.96
		2011	20	3.69	1.25	0.96
		2013	20	2.62	0.20	0.98
	8	2009	17	2.92	1.12	0.95
		2011	20	2.20	0.63	0.97
		2013	20	3.62	1.67	0.93
Math	4	2009	17	6.09	4.10	0.93
		2011	20	4.97	2.60	0.94
		2013	20	3.60	1.46	0.95
	8	2009	14	5.21	3.40	0.95
		2011	17	3.77	2.09	0.96
		2013	14	4.54	1.47	0.94
Average		2009	14-17	4.70	2.69	0.95
		2011	17-20	3.79	1.64	0.96
		2013	14-20	3.66	1.20	0.95
		Reading	17-20	3.23	1.17	0.96
		Math	14-20	4.77	2.52	0.95
		All	14-20	4.07	1.84	0.95
Subgroup Average		Male	14-20	4.14	1.84	0.97
		Female	14-20	3.95	1.70	0.98
		White	11-19	3.89	0.66	0.98
		Black	13-19	4.11	1.80	0.96
		Hispanic	12-20	4.07	2.08	0.94

Source: Authors calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data. Estimates are based on Equation 7 in text. Subgroup averages are computed from a model that pools across grades and years within subject (like Equation 9 in text); the subject averages are then pooled within subgroup.

Table 2: Precision-adjusted correlations of linked district means 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	1139	0.90	0.95
		2011	1472	0.87	0.93
		2013	1843	0.92	0.95
	8	2009	959	0.84	0.91
		2011	1273	0.87	0.91
		2013	1597	0.88	0.92
Math	4	2009	1128	0.86	0.93
		2011	1467	0.82	0.90
		2013	1841	0.87	0.93
	8	2009	970	0.83	0.93
		2011	1279	0.84	0.92
		2013	1545	0.87	0.95
Average		2009	4196	0.86	0.93
		2011	5491	0.85	0.91
		2013	6826	0.88	0.94
All Years			16513	0.87	0.93

Source: Authors' calculations from ED*Facts* and NWEA MAP data. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Sample includes districts with reported NWEA MAP scores for at least 90% of students.

Table 3. Estimated comparison of linked and TUDA district means, pooled across grades and years, by subject.

	Reading		Math	
Linked ED*Facts* Parameters				
Intercept (γ_{00})	228.53	***	250.59	***
	(2.00)		(2.10)	
Year (γ_{01})	0.91	***	0.43	*
	(0.17)		(0.19)	
Grade (γ_{02})	10.81	***	9.58	***
	(0.27)		(0.29)	
TUDA Parameters				
Intercept (γ_{10})	227.41	***	248.14	***
	(2.12)		(2.49)	
Year (γ_{11})	1.03	***	0.91	***
	(0.10)		(0.11)	
Grade (γ_{12})	10.84	***	9.68	***
	(0.22)		(0.17)	
L2 Intercept Variance - Linked (σ_0^2)	2.51		2.66	
L2 Intercept Variance - TUDA (σ_1^2)	0.82		1.26	
Correlation - L2 Residuals	1.00		0.36	
L3 Intercept Variance - Linked (τ_1^2)	8.87		9.27	
L3 Intercept Variance - TUDA (τ_4^2)	9.79		11.10	
L3 Year Slope Variance - Linked (τ_2^2)	--		--	
L3 Year Slope Variance - TUDA (τ_5^2)	--		--	
L3 Grade Slope Variance - Linked (τ_3^2)	1.06		1.03	
L3 Grade Slope Variance - TUDA (τ_6^2)	0.93		0.61	
Correlation - L3 Intercepts	0.98		0.98	
Correlation - L3 Year Slopes	--		--	
Correlation - L3 Grade Slopes	0.85		0.98	
Reliability L3 Intercept - Linked	0.98		0.98	
Reliability L3 Year Slope - Linked	--		--	
Reliability L3 Grade Slope - Linked	0.76		0.73	
Reliability L3 Intercept - TUDA	1.00		1.00	
Reliability L3 Year Slope - TUDA	--		--	
Reliability L3 Grade Slope - TUDA	0.87		0.72	
N – Observations	228		204	
N - Districts	20		20	

Source: Authors' calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data. Estimates are based on Equation 9 in text. *Note*: the level 3 random errors on the year slope were not statistically significant, and so were dropped from the model. L2 = "Level 2"; L3 = "Level 3".

Table 4. 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, measurement error adjusted.

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2011	20	3.78	0.89	0.95
	8	2011	20	2.14	1.47	0.99
Math	4	2011	20	4.67	2.25	0.94
	8	2011	14	3.81	1.66	0.96
Average		Reading	20	3.07	1.18	0.97
		Math	14-20	4.26	1.95	0.95
		All	14-20	3.72	1.57	0.96

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

Table 5: 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**	*0.93*	**0.93**	*0.94*
	4	0.95	*0.94*	0.93	*0.94*	0.95
	5	*0.94*	**0.94**	*0.93*	**0.93**	*0.94*
	6	*0.92*	**0.94**	*0.92*	**0.93**	*0.93*
	7	*0.92*	**0.93**	*0.92*	**0.92**	*0.92*
	8	0.91	*0.91*	0.91	*0.91*	0.92
Math	3	0.91	**0.89**	*0.91*	**0.91**	*0.91*
	4	0.93	*0.92*	0.90	*0.92*	0.93
	5	*0.91*	**0.90**	*0.92*	**0.91**	*0.93*
	6	*0.93*	**0.93**	*0.94*	**0.93**	*0.94*
	7	*0.94*	**0.95**	*0.95*	**0.95**	*0.95*
	8	0.93	*0.93*	0.92	*0.94*	0.95

No interpolation	0.93
Single interpolation	*0.93*
Double interpolation	**0.93**

Note. Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Sample includes districts with reported NWEA MAP scores for at least 90% of students.

Table 6: Reported and interpolated means and standard deviations for the National Assessment of Educational Progress, National Public School estimates, 2008-2013.

		Reading					
		2008	2009	2010	2011	2012	2013
Means	8	259.10	260.10	260.90	261.70	263.30	264.80
	7	248.50	249.30	250.00	250.70	252.10	253.40
	6	237.90	238.60	239.20	239.80	240.90	242.00
	5	227.30	227.80	228.30	228.80	229.70	230.50
	4	216.70	217.00	217.40	217.80	218.50	219.10
	3	206.10	206.20	206.50	206.80	207.30	207.70
SDs	8	36.80	36.30	36.10	35.80	35.60	35.30
	7	37.20	36.70	36.50	36.30	36.20	36.10
	6	37.50	37.00	36.90	36.90	36.90	36.90
	5	37.90	37.40	37.40	37.40	37.50	37.60
	4	38.20	37.70	37.80	37.90	38.20	38.40
	3	38.60	38.10	38.20	38.40	38.80	39.20

		Mathematics					
		2008	2009	2010	2011	2012	2013
Means	8	279.10	280.10	280.80	281.40	282.10	282.70
	7	268.80	269.60	270.20	270.90	271.50	272.10
	6	258.50	259.10	259.70	260.30	260.90	261.60
	5	248.20	248.60	249.20	249.80	250.40	251.00
	4	238.00	238.10	238.70	239.20	239.80	240.40
	3	227.70	227.60	228.10	228.70	229.20	229.80
SDs	8	37.70	37.60	37.40	37.10	37.10	37.10
	7	35.70	35.70	35.50	35.30	35.30	35.40
	6	33.80	33.70	33.60	33.40	33.60	33.70
	5	31.80	31.80	31.70	31.60	31.80	32.00
	4	29.90	29.80	29.80	29.70	30.00	30.30
	3	27.90	27.90	27.90	27.90	28.20	28.60

Note: Reported in 2009, 2011, and 2013 in grades 4 and 8, interpolated and extrapolated elsewhere. Lighter shaded cells are the basis for year-based scaling; darker shaded cells are the basis for cohort-based scaling. These are expanded population estimates and may differ slightly from those reported in public reports.

Table 7. Estimated parameters and variance components, model pooling estimated district means across grades and cohorts, by scale and subject.

	c* Scale				g* Scale			
	Mathematics		Reading		Mathematics		Reading	
Intercept (γ_{00})	-0.012	***	-0.006	*	5.465	***	5.457	***
	(0.003)		(0.003)		(0.010)		(0.010)	
Cohort (γ_{01})	0.017	***	0.016	***	0.049	***	0.049	***
	(0.000)		(0.000)		(0.001)		(0.001)	
Grade (γ_{02})	0.005	***	0.002	***	1.001	***	0.989	***
	(0.001)		(0.000)		(0.002)		(0.001)	
L2 Intercept Variance (σ^2)	0.016	***	0.007	***	0.135	***	0.066	***
L3 Intercept Variance (τ_0^2)	0.157	***	0.140	***	1.344	***	1.338	***
L3 Cohort Slope Variance (τ_1^2)	0.002	***	0.001	***	0.015	***	0.009	***
L3 Grade Slope Variance (τ_2^2)	0.004	***	0.002	***	0.037	***	0.020	***
Between District Proportion of Residual Variance ($\tau_0^2/(\tau_0^2+\sigma^2)$)	0.908		0.951		0.909		0.953	
Reliability of L3 Intercept	0.968		0.970		0.969		0.970	
Reliability of L3 Cohort Slope	0.701		0.655		0.698		0.659	
Reliability of L3 Grade Slope	0.727		0.693		0.755		0.696	
N - Observations	347131		356860		347131		356860	
N - Districts	14012		14030		14012		14030	

Note: Estimates are based on the model described by Equation (16) in text, fitted to estimated district means for grades 3-8 in 2009-2013. Source: Authors' calculations from ED*Facts* data.

Figure 1. Illustration of linear linking method

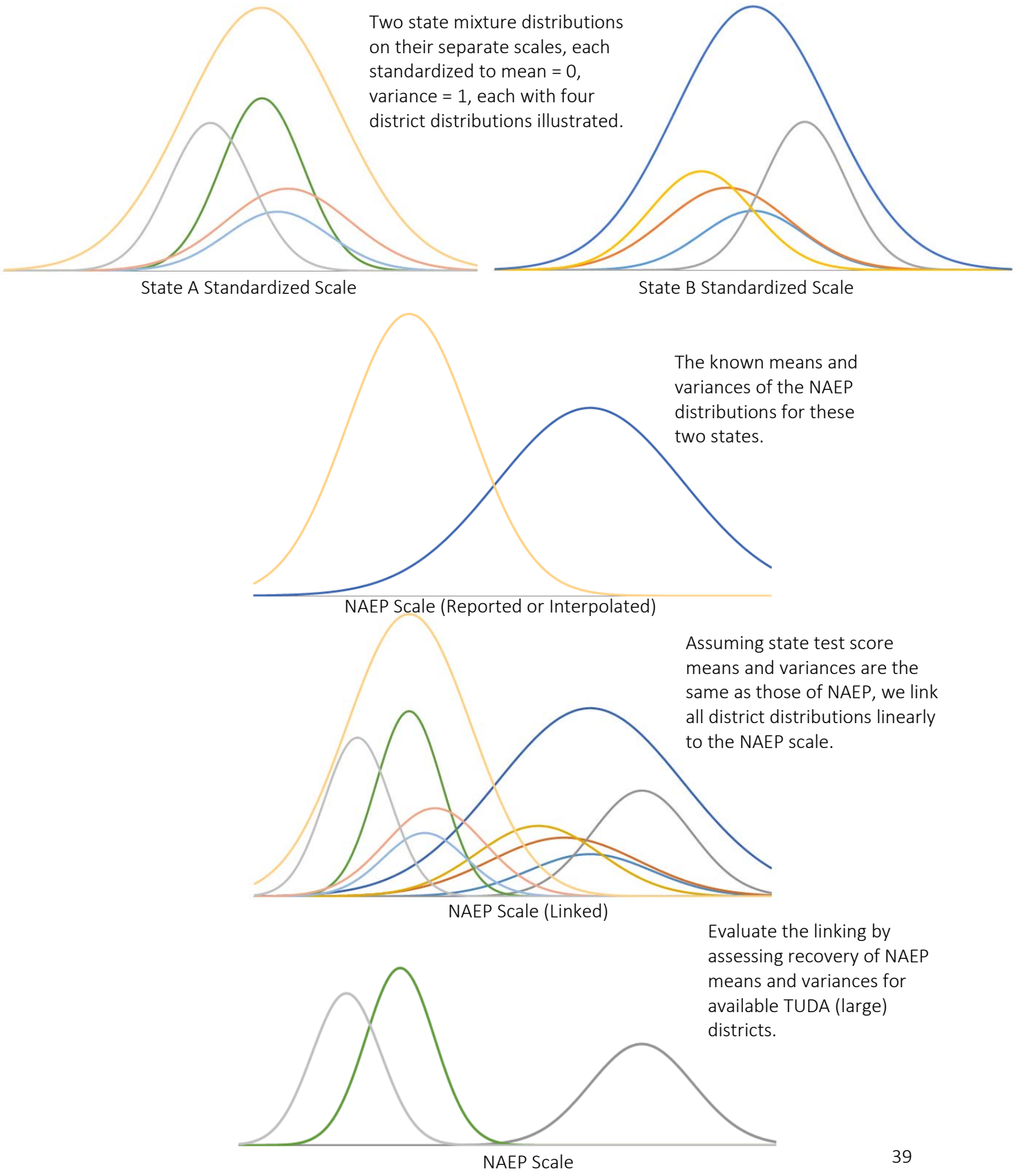
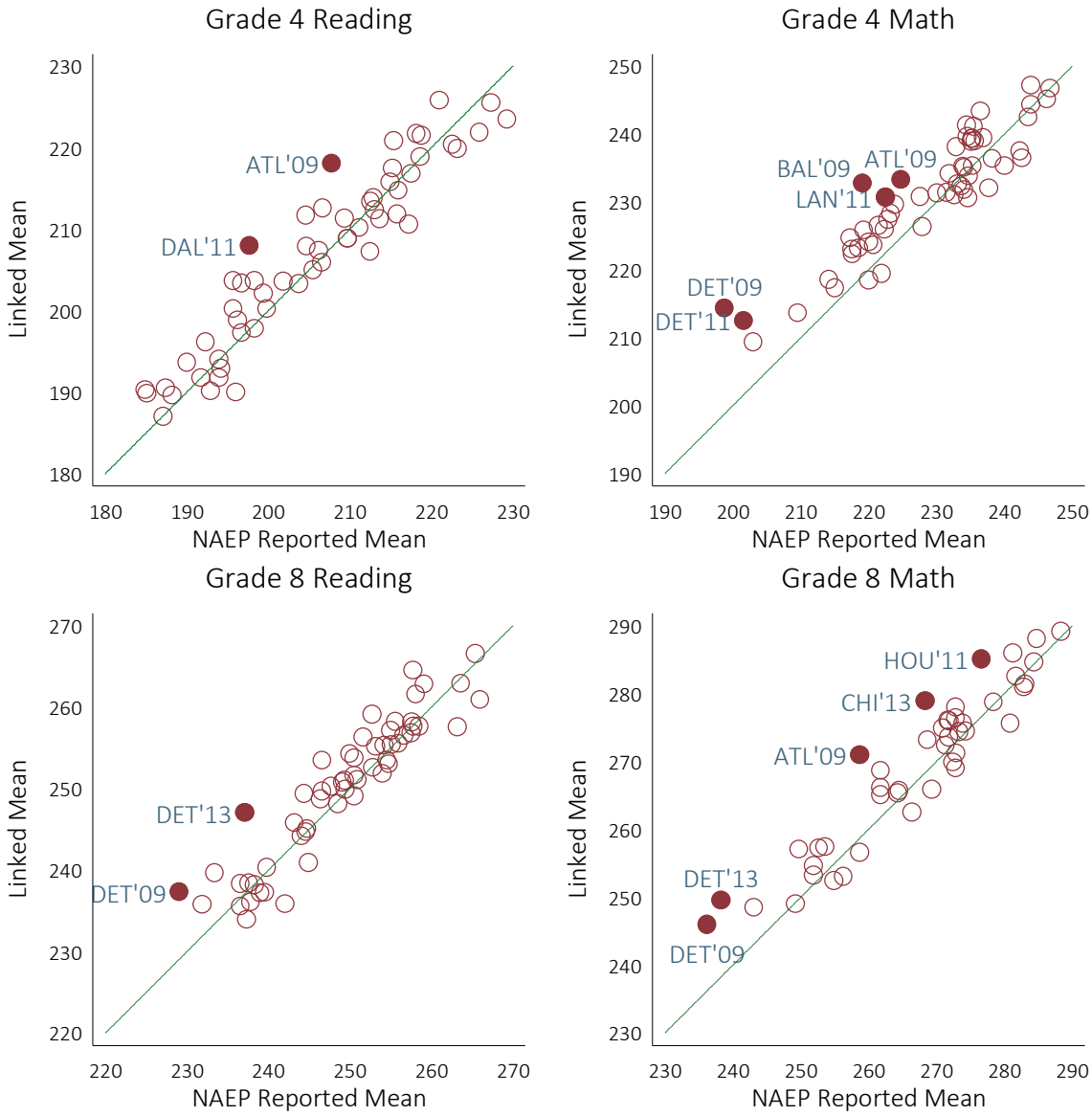
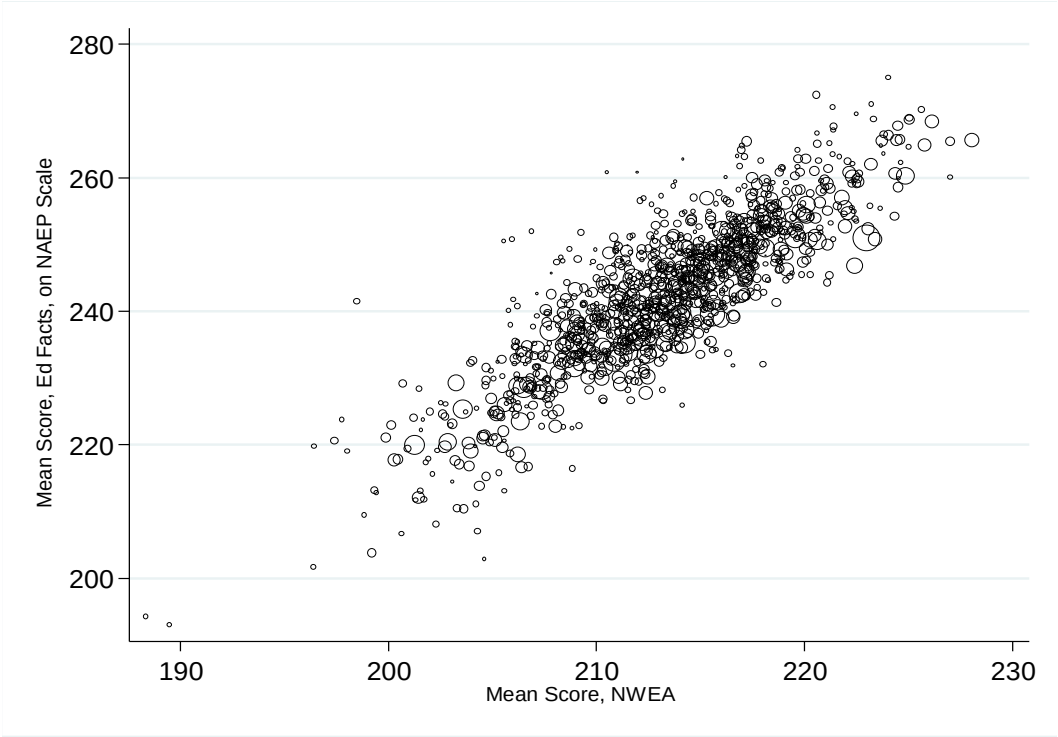


Figure 2: Comparison of reported means from NAEP TUDA and NAEP-linked state test score distributions, grades 4 and 8, Reading and Mathematics, in 2009, 2011, and 2013.



Note: DAL = Dallas; ATL = Atlanta; DET = Detroit; BAL = Baltimore; LAN = Los Angeles; CHI = Chicago; HOU = Houston. District-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.

Appendix Tables

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

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2009	17	0.89	-1.31	0.76
		2011	20	2.28	-0.14	0.46
		2013	20	1.07	0.34	0.96
	8	2009	17	1.84	-1.33	0.88
		2011	20	1.80	-1.00	0.41
		2013	20	1.01	-1.27	0.62
Math	4	2009	17	1.40	-0.09	0.72
		2011	20	1.71	0.83	0.68
		2013	20	1.64	0.99	0.71
	8	2009	14	2.15	-0.62	0.77
		2011	17	1.86	0.24	0.79
		2013	14	2.49	0.15	0.56
Average		2009	14-17	1.64	-0.84	0.78
		2011	17-20	1.93	-0.02	0.59
		2013	14-20	1.66	0.05	0.71
		Reading	17-20	1.57	-0.79	0.68
		Math	14-20	1.91	0.25	0.71
		All	14-20	1.75	-0.27	0.69
Subgroup Average		Male	14-20	1.60	-0.13	0.72
		Female	14-20	1.76	0.05	0.65
		White	11-19	4.43	2.99	0.42
		Black	13-19	1.90	-0.24	0.40
		Hispanic	12-20	2.51	-1.24	0.34

Source: Authors calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data. Estimates are based on Equation 7 in text. Subgroup averages are computed from a model that pools across grades and years within subject (like Equation 9 in text); the subject averages are then pooled within subgroup.

Table A2: Precision-adjusted correlations of linked district standard deviations with NWEA MAP district standard deviations 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	1139	0.52	0.58
		2011	1472	0.55	0.61
		2013	1843	0.59	0.64
	8	2009	959	0.54	0.57
		2011	1273	0.51	0.58
		2013	1597	0.52	0.51
Math	4	2009	1128	0.65	0.75
		2011	1467	0.58	0.66
		2013	1841	0.63	0.69
	8	2009	970	0.65	0.70
		2011	1279	0.55	0.62
		2013	1545	0.65	0.70
Average		2009	4196	0.59	0.65
		2011	5491	0.55	0.62
		2013	6826	0.60	0.64
All Years			16513	0.58	0.63

Source: Authors' calculations from ED*Facts* and NWEA MAP data. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Sample includes districts with reported NWEA MAP scores for at least 90% of students.

Table A3. Estimated comparison of linked and TUDA district standard deviations, pooled across grades and years, by subject.

	Reading		Math	
Linked ED*Facts* Parameters				
Intercept (γ_{00})	36.21	***	33.17	***
	(0.42)		(0.35)	
Year (γ_{01})	0.18	+	0.39	***
	(0.10)		(0.11)	
Grade (γ_{02})	-0.68	***	1.66	***
	(0.13)		(0.18)	
TUDA Parameters				
Intercept (γ_{10})	36.85	***	32.75	***
	(0.39)		(0.36)	
Year (γ_{11})	0.07		0.17	*
	(0.10)		(0.08)	
Grade (γ_{12})	-0.38	**	1.81	***
	(0.13)		(0.14)	
L2 Intercept Variance - Linked (σ_0^2)	0.96		1.21	
L2 Intercept Variance - TUDA (σ_1^2)	0.74		0.43	
Correlation - L2 Residuals	1.00		1.00	
L3 Intercept Variance - Linked (τ_1^2)	1.73		1.30	
L3 Intercept Variance - TUDA (τ_4^2)	1.60		1.43	
L3 Year Slope Variance - Linked (τ_2^2)	--		--	
L3 Year Slope Variance - TUDA (τ_5^2)	--		--	
L3 Grade Slope Variance - Linked (τ_3^2)	0.44		0.67	
L3 Grade Slope Variance - TUDA (τ_6^2)	0.46		0.54	
Correlation - L3 Intercepts	0.60		0.70	
Correlation - L3 Year Slopes	--		--	
Correlation - L3 Grade Slopes	0.77		0.89	
Reliability L3 Intercept - Linked	0.86		0.76	
Reliability L3 Year Slope - Linked	--		--	
Reliability L3 Grade Slope - Linked	0.63		0.77	
Reliability L3 Intercept - TUDA	0.84		0.87	
Reliability L3 Year Slope - TUDA	--		--	
Reliability L3 Grade Slope - TUDA	0.63		0.79	
N - Observations	228		204	
N - Districts	20		20	

Source: Authors' calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data. Estimates are based on Equation 9 in text. *Note*: the level 3 random errors on the year slope were not statistically significant, and so were dropped from the model. L2 = "Level 2"; L3 = "Level 3".

Table A4. 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, measurement error adjusted.

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2011	20	2.12	0.39	0.53
	8	2011	20	2.53	-0.70	0.00
Math	4	2011	20	2.15	1.43	0.53
	8	2011	14	1.52	0.65	0.93
Average		Reading	20	2.34	-0.15	0.26
		Math	14-20	1.86	1.04	0.73
		All	14-20	2.11	0.44	0.50

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

Table A5: Precision-adjusted correlations between NWEA MAP district standard deviations and NAEP-linked estimates.

Subject	Grade	2009	2010	2011	2012	2013
Reading	3	*0.53*	0.57	*0.63*	0.62	*0.61*
	4	0.58	*0.57*	0.61	*0.62*	0.64
	5	*0.61*	0.54	*0.58*	0.64	*0.64*
	6	*0.60*	0.58	*0.61*	0.64	*0.57*
	7	*0.60*	0.57	*0.56*	0.55	*0.53*
	8	0.57	*0.56*	0.58	*0.52*	0.51
Math	3	0.71	0.70	*0.69*	0.67	*0.65*
	4	0.75	*0.77*	0.66	*0.71*	0.69
	5	*0.73*	0.69	*0.73*	0.72	*0.74*
	6	*0.73*	0.75	*0.69*	0.71	*0.73*
	7	*0.75*	0.65	*0.70*	0.71	*0.70*
	8	0.70	*0.64*	0.62	*0.71*	0.70

No interpolation	0.63
Single interpolation	*0.58*
Double interpolation	0.64

Note. Linked using NAEP Expanded Population Estimates. NWEA MAP = Northwest Evaluation Association Measures of Academic Progress. Sample includes districts with reported NWEA MAP scores for at least 90% of students.

Table A6. Estimated parameters and variance components, model pooling estimated district standard deviations across grades and cohorts, by scale and subject.

	c* Scale				g* Scale			
	Mathematics		Reading		Mathematics		Reading	
Intercept (γ_{00})	0.826	***	0.863	***	2.409	***	2.671	***
	(0.001)		(0.001)		(0.002)		(0.002)	
Cohort (γ_{01})	0.001	***	0.000		0.003	***	0.000	
	(0.000)		(0.000)		(0.000)		(0.000)	
Grade (γ_{02})	-0.004	***	-0.001	***	0.125	***	-0.046	***
	(0.000)		(0.000)		(0.001)		(0.001)	
L2 Intercept Variance (σ^2)	0.001	***	0.001	***	0.008	***	0.005	***
L3 Intercept Variance (τ_0^2)	0.006	***	0.005	***	0.050	***	0.044	***
L3 Cohort Slope Variance (τ_1^2)	0.000	***	0.000	***	0.001	***	0.001	***
L3 Grade Slope Variance (τ_2^2)	0.000	***	0.000	***	0.002	***	0.002	***
Between District Proportion of Residual Variance ($\tau_0^2/(\tau_0^2+\sigma^2)$)	0.867		0.889		0.866		0.892	
Reliability of L3 Intercept	0.935		0.931		0.934		0.932	
Reliability of L3 Cohort Slope	0.594		0.562		0.588		0.564	
Reliability of L3 Grade Slope	0.587		0.575		0.587		0.583	
N - Observations	347131		356860		347131		356860	
N - Districts	14012		14030		14012		14030	

Note: Estimates are based on the model described by Equation (16) in text, fitted to estimated district standard deviations for grades 3-8 in 2009-2013. Source: Authors' calculations from ED*Facts* data.