Two Watches:

Measurement Error Models for Estimating Educational Progress from Multiple Test Score Trends

sean f. reardon
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

Jie Min
Stanford University

Version August 20, 2026

This paper benefitted from feedback at the annual meeting of the National Council on Measurement in Education, including fellow panelists Scott Marion, Christine Rozunick, and Peggy Carr; from participants at the Applied Statistics Workshop at the Institute for Quantitative Social Sciences at Harvard University and the Research, Evaluation, Measurement, and Policy program at the University of Massachusetts Amherst; and from fellow members of the Educational Opportunity Project, including Erin Fahle, Benjamin Shear, jim saliba, Jiyeon Shim, and Demetra Kalogrides. Emily Oster and Clare Halloran at the Education Data Center and staff at the National Center for Education Statistics graciously provided essential data for the paper. This paper was supported by funding from the Gates Foundation. The opinions expressed here are ours and do not reflect the views of the funders or data providers.

Two Watches:

Measurement Error Models for Estimating Educational Progress

from Multiple Test Score Trends

Abstract

Understanding large-scale educational progress often requires reconciling information from multiple testing programs that differ in their purpose, precision, and periodicity. Like watches that disagree about the time, tests may report different trends for the same populations, subjects, and periods. We develop a precision-adjusted multilevel measurement error model of the relationship between 2-year National Assessment of Educational Progress (NAEP) score trends and the corresponding state test score trends. Using trend data from NAEP and state testing programs, the model jointly estimates their true variances and correlation, their reliabilities, and systematic discrepancies between them.

We find that sampling error in NAEP and estimated linking error in state assessments lead to moderately low reliability of trend estimates (reliabilities of 0.45 for NAEP and between 0.49 and 0.61 for state tests). We estimate the correlation between 2-year NAEP and state test trends is between 0.74 and 0.83. On average, state test trends overestimate NAEP trends by approximately 0.05 standard deviations units per two-year period. We demonstrate that Empirical Bayes estimators based on the model may produce substantially more accurate estimates of educational progress by combining information from NAEP and state assessments or, when one is missing, predicting it from the other.

Introduction

Different educational tests may report discrepant trends for the same populations, subjects, and time periods. For example, state testing programs and the National Assessment of Educational Progress (NAEP) can report different trends in achievement for the same state, grade, subject, and years. Similarly, districts may need to reconcile trends from state tests and district-administered benchmark assessments. This is reminiscent of the measurement aphorism, "a person with one watch knows what time it is; a person with two watches is never quite sure" (e.g., Brennan, 2001, p. 295). Rather than picking one of two "watches" or simply averaging them, we develop a modeling framework that quantifies their discrepancies and enables an optimal estimate of trends when either or both are available.

Previous research has documented substantial discrepancies between NAEP and state test score trends (Fuller et al., 2007; Ho & Polikoff, 2025; Koretz, 2008; Koretz et al., 2001).[1] Two studies pool observed discrepancies across multiple states: Ho (2007) finds that two-year state test trends exceeded two-year NAEP trends by an average of 0.08 standard deviation (SD) units across 26 states from 2003-2005; Jacob (2007) finds the same overall magnitude for two-year trends across 4 states in the 1990s.

These discrepancies have multiple potential explanations. Some may reflect meaningful differences in student progress on what tests respectively measure: different constructs, complexity, or curricula (e.g., Polikoff, 2012). Others may arise from validity threats to state test score trends, including teaching to the test, strategic teaching to subpopulations, strategic selection of students or teachers, or outright cheating (Booher-Jennings, 2005; Ho & Polikoff, 2025; Koretz, 2008; Neal & Schanzenbach, 2010; State of Georgia, 2011). Trends can also differ due to imprecision, an issue we address in this paper.

[1] An indirect indication of State-NAEP trend discrepancies is drift in mapped proficiency cut scores when neither state tests nor their achievement levels change. If a mapped state proficiency standard appears to decrease on the NAEP scale, this is because the implied state trend exceeds the NAEP trend in that region of the score scale (Ho & Haertel, 2007). These analyses are conducted regularly (e.g., Bandeira de Mello et al., 2009; Braun & Qian, 2007; Ji et al., 2021); however, their goal is to map standards rather than interpret drift as discrepant trends. In this paper, we estimate the trend discrepancy directly using all available cut score data, not just the proficiency cut score.

Table 1 reviews general differences between NAEP and state tests along 10 dimensions. These are summarized well by prior literature referenced above. Three issues are particularly relevant for our research questions and modeling approach. First, theories of test score inflation (Koretz, 2008; Koretz & Hamilton, 2006) suggest that trends should be more positive for state tests than for NAEP, particularly in the early years of a new testing program (Table 1, #2, #8). Second, near-census testing for state tests, compared to smaller samples for state NAEP, predicts greater precision for state test score trends (Table 1, #3). Third, states change content and performance standards periodically, so state test score trends are not always available for matched NAEP years (Table 1, #6, #7).

Table 1. Ten general contrasts between state tests and NAEP between 2009 and 2024

Contrast	State Tests	NAEP
1. Primary purpose(s)	School accountability; student-level reporting	Population monitoring (aggregate progress)
2. Stakes / incentives	Higher stakes for schools. Stakes for teachers and students in some jurisdictions/years	Lower stakes for states and large districts (“public accountability”)
3. Sampling	Near-census (≈95–99%)*	Sample, typically ~1750 students in ~100 schools for each state-subject-grade.
4. Periodicity / timing	Annual (typically spring)	Biennial (‘09, ‘11, ‘13, ‘15, ‘17, ‘19, ‘22, ‘24. Spring.)
5. Reporting levels	State; district; school; student.	States; Large Urban Districts
6. Comparability	Typically across years, not across states	Across years, states, and large districts
7. Trend breakage	Common at irregular intervals.	Rare and avoided.
8. Construct / Content	State-determined content standards.	Consensus frameworks via a governing board.
9. Test design / form	Largely fixed forms, sometimes adaptive	Matrix sampling; plausible values.
10. Grades / subjects	Grades 3-8 + 1 high school, English Language Arts and Math	State-level results in grades 4 and 8, Reading and Math

*Exceptions: alternate assessments, severe cognitive disabilities, opt-out, absences/invalidations, etc.

Prior research on discrepancies between state and NAEP trends has been limited in scope to selected years and states and did not account for imprecision in state and NAEP trends. Student-level measurement error in state test scores attenuates the magnitude of trends when scores are standardized to the observed test score distribution. Sampling error and other forms of error in state test score means attenuate trend correlations. We develop a precision-adjusted multilevel measurement error model of the relationship between NAEP and state test score trends from 2009 to 2024. The model estimates true (error-corrected) variance and correlations, given known trend measurement error variances. The model also implies circumstances when using both state test and NAEP trends together can improve estimation of true NAEP trends. In a patchwork landscape of educational test score trends, this modeling framework and its estimates can improve understanding of aggregate educational progress.

Trend estimates carry sources of error beyond those that single-timepoint error estimates imply. Although single-timepoint error variances for NAEP and state test means and standard deviations are available (Mislevy et al., 1992; Reardon et al., 2017), trends are susceptible to additional imprecision in the link between year-to-year score scales. We call "equating error" the imprecision introduced by the process of equating scores using a finite number of common items embedded across test forms over time (Kolen & Brennan, 2014; Michaelides & Haertel, 2014). This source affects both NAEP and state test trends; however, we follow NAEP program documentation that suggests this component is negligible relative to NAEP sampling error (B. Cramer, personal communication, May 2026; see also Mazzeo et al., 2025). We call "discretization error" the imprecision that arises from the mapping of scale scores to discrete score categories that can then fall on either side of state proficiency cut score standards (Ho & Yu, 2015; Yee & Ho, 2015). This source affects only our estimation of state trends.[2] We refer to the sum of equating error and discretization error as "linking error" for state test score trends.

[2] We estimate state test score means and standard deviations using heteroskedastic ordered probit (HETOP) models fit to categorical data from state-defined proficiency categories (Reardon et al., 2017). This accounts for student sampling error and the imprecision arising from coarsening scores into a small number of proficiency categories.

Because states and their vendors report neither equating nor discretization error variances, we organize our analysis by what these unmodeled errors do and do not affect. As we will demonstrate, the reliability of NAEP trends, the bias of state trends, and each estimator of true NAEP change and its mean squared error depend on the observed variance (the sum of true and error variances) of state trends. Unmodeled state trend error repartitions this sum but leaves the total observed variance unchanged. State trend reliability and the true correlation between state and NAEP trends, in contrast, depend on the partitioning of this error and therefore on the magnitude of linking error. We address two sets of research questions accordingly. First, under the working assumption that there is no additional linking error in state test score trends, we ask:

1. What are the reliabilities of NAEP and state test score trends, and what is the systematic bias of one as an estimate of the other?
2. What are the biases and mean squared errors of five estimators of true NAEP trends: i) observed NAEP trends, ii) observed state trends, iii) bias-corrected state trends, iv) Empirical Bayes (shrinkage) estimates based on observed state trends, and v) Empirical Bayes estimates based on both observed state and NAEP trends? How well can we expect to predict NAEP trends where both tests are available versus where only state test data may exist?

Given that the assumption of no linking error in state test score trends is implausible, we next conduct a bounding exercise to estimate state linking error from the within-state variance of trends across grades, net of the true grade-trend variance we recover from NAEP. We use this estimated linking error variance to estimate several quantities that depend on unmodeled state-trend error:

However, additional error from this coarsening interacts with equating error, as discrete bins with large percentages of students can fall on either side of cut score boundaries (Yee & Ho, 2015). As noted, we call this error that affects trend estimates “discretization error” and distinguish it from “equating error” affecting trends due to item sampling.

3. Accounting for state test linking error, what is the reliability of state test trends, and what is the true correlation between state and NAEP trends?

Consistent with prior research, we find that average state test score trends are highly correlated with, but more positive than, NAEP trends for available state trend eras between 2009 and 2024. We estimate the correlation of 2-year state test and NAEP trends is between 0.74 and 0.83, depending on the magnitude of linking error in estimated state test trends. Moreover, we estimate that two-year state test trends overestimate true NAEP trends, on average, by 0.05 standard deviation units (equivalently, two-year NAEP trends underestimate state test trends by the same amount). Despite this discrepancy (bias), in some cases, Empirical Bayes (shrunken) estimates using state test score trends can be better estimates of true NAEP trends than observed NAEP trends themselves, as measured by root mean squared error (RMSE). This result emerges because the higher reliability of observed state test changes (relative to observed NAEP changes) and their high true correlation with NAEP trends outweigh the disadvantages of their systematic bias and imperfect alignment with NAEP.

Our model is a tool for mapping the fragmented landscape of large-scale educational progress in the United States. NAEP provides unbiased but imprecise estimates at biennial intervals. State tests provide somewhat more precise and more frequent (annual) estimates, but with systematic positive bias and frequent trend breaks that prevent long-term monitoring. From this model, we derive an optimal trend estimator, combining information from both state and NAEP tests when state tests are stable over time, improving inference about true educational progress. While often broken and discrepant from NAEP trends, state trends can improve estimation of educational progress. Our results contribute both to methods for analyzing trend data and to policy options for using multiple measures in educational monitoring systems.

Data

To estimate test score trends, we use data from the NAEP Data Explorer, the ED*Facts* Initiative, and data collected from state assessment programs by the Education Data Center (EDC). For NAEP trends, we use state means and standard deviations, as well as their standard errors, from 2009-2024. We include 50 states and Washington DC, in grades 4 and 8, in reading and mathematics. For state test trends, we use 2009-2019 data from ED*Facts*; and 2022-2024 data from EDC (2025). To estimate valid state test score trends, state test score scales and achievement levels must remain comparable from one year to the next. To evaluate comparability, we use EDC data (2025). The EDC research team reviews a range of data sources to evaluate whether cut scores are comparable over time, including technical manuals, press releases, stakeholder presentations, interpretation guides, state board meeting materials, and discussions with state agency representatives. We use the EDC "test change" variable to exclude trends that confound changes in cut score meaning with changes in proficiency.

The EDC "test change" variable is available for all states from 2024 retrospectively through 2019, but availability becomes sparser in earlier years. Roughly 2/3 of states have availability retrospectively to 2015, but only a few states have availability retrospectively to 2009. To determine state test comparability where EDC is unavailable, we review a similar range of data sources as EDC through archival material online. We complement these with ED*Facts* surveys of state agency representatives about test changes (e.g., U.S. Department of Education, 2016).[3]

[3] To maximize the chance of accurately flagging broken state trends, we also conducted a preliminary analysis comparing state and NAEP trend magnitudes. We flagged state-NAEP trend discrepancies greater than 0.15 SD units for manual review of test comparability documentation. To guard against confirmation bias, where we might be more likely to find documentation of test changes simply because a trend looks anomalous, we embedded a set of placebo cases with small discrepancies (less than 0.05 SD units) into the review list. We examined all cases without knowing which were flagged as large discrepancies and which were placebos. We identified a small number of cases where historical documentation revealed true changes. None of these were placebos. The concentration of identified test changes among flagged cases, and their absence among placebos, provides evidence that our review process was sensitive to documentation of test changes and not to the size of the trend discrepancy.

To estimate trends from state testing programs when trends are comparable, we use population-level counts of students in ordered proficiency categories from the ED*Facts* initiative. Aligning with NAEP, we include 50 states and Washington DC, in grades 4 and 8, in Reading or English Language Arts (RLA) and mathematics. We fit heteroskedastic ordered probit (HETOP) models to these state proficiency counts, yielding estimated state means and variances and their standard errors (Reardon et al., 2017). We standardize both NAEP and state test trends to the first year of each consecutive year-pair, so each test score trend magnitude is interpretable as a mean difference in standard deviation (SD) units of the first year of the state-year-pair. Following Reardon et al. (2021), we use state-reported reliability estimates to disattenuate state effect sizes, because state standard deviations are inflated slightly by measurement error.[4]

Table 2. Comparable state test score scales and estimable trends by trend years (of 204 possible state-subject-grade combinations per row).

	Comparable Scales		Estimable Trends	
Trend Years	n	%	n	%
2009-2011	154	75.5%	133	65.2%
2011-2013	153	75.0%	142	69.6%
2013-2015	34	16.7%	26	12.7%
2015-2017	140	68.6%	105	51.5%
2017-2019	112	54.9%	93	45.6%
2019-2022	160	78.4%	122	59.8%
2022-2024	156	76.5%	129	63.2%
All	909	63.7%	750	52.5%

Table 2 shows the number and percent of comparable scales and estimable state trends per trend year. There are 204 possible trends per year (50 states plus Washington DC, grades 4 and 8, in reading or

[4] Measurement error in state test scores will increase observed standard deviations and deflate trend magnitudes expressed in standard deviation units. We gather reliability estimates from state technical manuals and impute reliability estimates when these are not available. As we note in Reardon et al. (2021), the reported reliabilities are extremely consistent across states, grades, years, and subjects, suggesting that imputation is not consequential. We do not adjust NAEP standard deviations, as NAEP estimation procedures account for measurement error due to item assignment to examinees (Mislevy et al., 1992).

English Language Arts and mathematics). Comparable trends are those with no reported test score scale or cut scores change in the two-year period. Table 2 shows that state test score trends "break" at moderate frequency. Roughly a quarter of state trends "broke" in NAEP trend eras 2009-11 and 2011-13. As many states adopted new tests in the "Common Core" era (see Briggs, 2024, for a history), only 16.7% of state-subject-grade trends were comparable from 2013-2015. Since 2015, roughly two out of three state tests are unchanged through a matched NAEP era.

In some cases, in spite of comparable scales, EDFacts or EDC data were not available or sufficient to estimate trends using HETOP. This may occur because EDFacts has scores in only two categories (one proficiency cut score is insufficient to estimate changes in variance); because participation rates in state test scores are below 95%; or because a state administered end-of-course rather than end-of-year assessments (this is most common in 8th grade math), so that not all students in the state-grade-subject took the same test. All NAEP trends and all estimable state trends are the data to which we fit our modeling framework in the next section.

State and NAEP Trend Modeling Framework

Let δ_{ygbts} denote the true (latent) change in average scores from year $y-2$ to y in grade g and subject b on test t in state s. Let T indicate a state test and N indicate NAEP, so that δ_{ygbNs} is the true change in average NAEP scores between years $y-2$ and y; δ_{ygbTs} is the true change in average scores on the state test during the same period. We express changes in standard deviation units of the state-grade-subject-test specific year $y-2$ score distribution.

For simplicity of exposition, we focus first on changes between a single pair of years in a single grade and subject and drop the y, g, and b subscripts. Letting $\boldsymbol{\delta}_s$ denote the 2×1 vector of NAEP and state test changes, the joint distribution of the two changes is:

$$\boldsymbol{\delta}_s = \begin{bmatrix}\delta_{Ns} \\ \delta_{Ts}\end{bmatrix} \sim \left[\begin{pmatrix}\gamma_N \\ \gamma_T\end{pmatrix}, \begin{pmatrix}\tau_0 & \tau_{01} \\ \tau_{01} & \tau_1\end{pmatrix}\right] = [\boldsymbol{\Gamma}, \boldsymbol{\tau}]$$

(1)

If the state and NAEP tests measure the same underlying true change, then $\delta_{Ns} = \delta_{Ts}$ for all s, implying that $\gamma_N = \gamma_T$ and $\tau_0 = \tau_1 = \tau_{01}$. If $\gamma_N \neq \gamma_T$, then the change in average scores on the state test is a biased estimator of the true average change in NAEP scores, with bias $b = \gamma_T - \gamma_N$. If $\left|\frac{\tau_{01}}{\sqrt{\tau_0 \tau_1}}\right| < 1$, then the true NAEP and state test score changes are imperfectly correlated, implying the two tests measure different underlying sets of skills.

In practice, we observe an error-prone measure of $\boldsymbol{\delta}_s$: $\widehat{\boldsymbol{\delta}}_s = \boldsymbol{\delta}_s + \boldsymbol{e}_s$, where we assume the errors e_{Ns} and e_{Ts} are independent, mean-zero, with state- and test-specific variance:

$$\boldsymbol{e}_s \sim \left[\binom{0}{0}, \begin{pmatrix} \sigma_{Ns} & 0 \\ 0 & \sigma_{Ts} \end{pmatrix}\right] = (\mathbf{0}, \boldsymbol{\sigma}_s).$$

(2)

Under the assumption that $\mathbf{e}_s \perp \boldsymbol{\delta}_s$, the joint distribution of the true and observed NAEP and state changes is:

$$\begin{bmatrix} \boldsymbol{\delta}_s \\ \widehat{\boldsymbol{\delta}}_s \end{bmatrix} = \begin{bmatrix} \delta_{Ns} \\ \delta_{Ts} \\ \hat{\delta}_{Ns} \\ \hat{\delta}_{Ts} \end{bmatrix} \sim \left[\begin{pmatrix} \gamma_N \\ \gamma_T \\ \gamma_N \\ \gamma_T \end{pmatrix}, \begin{pmatrix} \tau_0 & \tau_{01} & \tau_0 & \tau_{01} \\ \tau_{01} & \tau_1 & \tau_{01} & \tau_1 \\ \tau_0 & \tau_{01} & \tau_0 + \sigma_{Ns} & \tau_{01} \\ \tau_{01} & \tau_1 & \tau_{01} & \tau_1 + \sigma_{Ts} \end{pmatrix}\right] = \left[\binom{\boldsymbol{\Gamma}}{\boldsymbol{\Gamma}}, \begin{pmatrix} \boldsymbol{\tau} & \boldsymbol{\tau} \\ \boldsymbol{\tau} & \boldsymbol{\tau} + \boldsymbol{\sigma}_s \end{pmatrix}\right]$$

(3)

This joint distribution implies the following data generating model:

$$\hat{\delta}_{ts} = \delta_{Ns}(N) + \delta_{Ts}(T) + e_{ts}$$

$$\delta_{Ns} = \gamma_N + u_{Ns}$$

$$\delta_{Ts} = \gamma_T + u_{Ts}$$

$$e_{ts} \sim (0, \sigma_{ts}); [u_{Ns}, u_{Ts}]' = \mathbf{u}_s \sim (\mathbf{0}, \boldsymbol{\tau})$$

(4a)

where N and T are dummy variables indicating whether $\hat{\delta}_{ts}$ is the observed change on the NAEP or state test, respectively. Or, more compactly,

$$\widehat{\boldsymbol{\delta}}_s = \boldsymbol{\Gamma} + \mathbf{u}_s + \mathbf{e}_s;$$

$$\mathbf{e}_s \sim (\mathbf{0}, \boldsymbol{\sigma}_s), \mathbf{u}_s \sim (\mathbf{0}, \boldsymbol{\tau})$$

(4b)

A. Estimating the bias and correlation of state and NAEP test score changes

Given the observed $\widehat{\boldsymbol{\delta}}_s$ and $\boldsymbol{\sigma}_s$, we can estimate $\boldsymbol{\Gamma}$ and $\boldsymbol{\tau}$, treating (4) as a multivariate "V-known" multilevel model or a multivariate meta-analytic model (Raudenbush & Bryk, 2002). We estimate the bias of the state test change as an estimate of the NAEP test change as $\hat{b} = \gamma_T - \gamma_N$ and the correlation of the true NAEP and state test score changes as $\hat{r} = \hat{\tau}_{01}/\sqrt{\hat{\tau}_0\hat{\tau}_1}$. We estimate $\hat{b}$ in two steps, described below.

We estimate model (4) from the observed state and NAEP test score changes in each pair of adjacent NAEP testing years from 2009-2019 and in 2022-2024. Estimates of NAEP trend variances and reliabilities draw from all NAEP trends, whereas state test trends and relationships between state trends and NAEP draw only from trends where states, subjects, grades, and consecutive year-pairs are available for both state tests and NAEP. We fit model (4) twice for each specification. We take $\boldsymbol{\tau}$ and γ_N from a model with all available NAEP data. We take $\hat{b} = \hat{\gamma}_T - \hat{\gamma}_N$ from a model fit using the paired sample, and calculate $\hat{\gamma}'_T = \hat{\gamma}_N + \hat{b}$.[5] We do not use data from the 2019-2022 changes both because test score changes were unusually large during the pandemic and because the bias and correlation of changes may be different over 3 years than over 2 years.

We fit model (4) to different subsets of the data, pooling in some cases over 4th and 8th grade and over math and reading test within year-pairs; in other cases, pooling over year-pairs within grade and subject; and in other cases pooling over all year-pairs, grades, and subjects. In models where we pool over grades, subjects, and/or year-pairs, we include a fully saturated set of grade-x-subject-x-year fixed effects, so that the estimates reflect the average within grade-subject-year changes and variances/covariances.

[5] This affects only the average state change, the bias, and the RMSE of the state as NAEP estimate. All other reported statistics are identical under either model.

Fitting model (4) requires that we know the state-specific error variances σ_{Ns} and σ_{Ts} of the NAEP and state test score changes. While we know the sampling error variance for NAEP score means and the total sampling error variance and HETOP estimation error variance of the state score means, we do not have precise estimates of the components of error variance due to equating error or discretization error in state tests. As we describe above, for NAEP, equating error is assumed present but negligible relative to sampling error given the robust matrix sampling design employed (Mislevy et al., 1992). Discretization error is also irrelevant for NAEP, since the NAEP means are estimated from scale scores rather than from proficiency category counts. But for state test means, linking error from both equating and discretization errors may be present.

To address this, we assume a range of values of additional error variance in state test means (above their known sampling and HETOP estimation error variance). In our initial models, we assume zero error variance from these sources, where HETOP standard errors capture all sources of error variance. This approach will lead to an overestimate of the reliability of state test score trends and an underestimate of the true correlation of NAEP and state trends. In subsequent analyses, we add additional error variance to the state test score estimates to identify a plausible range of estimated correlations.

B. Estimating true test score changes

Our second goal is to estimate $\boldsymbol{\delta}_s$ from $\widehat{\boldsymbol{\delta}}_s$. Here we discuss five possible estimators, and construct formulas for the bias and mean squared error of each, as well as for their correlation with $\boldsymbol{\delta}_s$. Given the observed error-prone NAEP and state trends, we have several ways of estimating the true trends δ_{Ns} and δ_{Ts} in state s.

First, and most obviously, $\hat{\delta}_{Ns}$ provides an estimate of δ_{Ns}. Given $e_{Ns}\sim(0,\sigma_{Ns})$, $\hat{\delta}_{Ns}$ will be an unbiased estimator of δ_{Ns} with MSE σ_{Ns}. Likewise $\hat{\delta}_{Ts}$ provides an unbiased estimate of δ_{Ts}, with MSE σ_{Ts}:

$$bias(\hat{\delta}_{Ns}) = E[\hat{\delta}_{Ns} - \delta_{Ns}] = 0;$$

$$MSE(\hat{\delta}_{Ns}) = E\left[(\hat{\delta}_{Ns} - \delta_{Ns})^2\right] = \sigma_{Ns}$$

$$bias(\hat{\delta}_{Ts}) = E[\hat{\delta}_{Ts} - \delta_{Ts}] = 0;$$

$$MSE(\hat{\delta}_{Ts}) = E\left[(\hat{\delta}_{Ts} - \delta_{Ts})^2\right] = \sigma_{Ts}$$

(5)

Second, each of the observed test score changes—$\hat{\delta}_{Ts}$ and $\hat{\delta}_{Ns}$—provides an estimate of the true change in scores on the other test. Intuitively, the observed trend in one test will provide a more useful estimate of the true trend in the other test to the extent that a) the true test trends are equal and b) the observed test trend is measured with little error. Given the joint distribution in (3), the bias and MSE of $\hat{\delta}_{Ts}$ as an estimator of δ_{Ns} and of $\hat{\delta}_{Ns}$ as an estimator of δ_{Ts} are:

$$bias(\hat{\delta}_{Ts}) = E[\hat{\delta}_{Ts} - \delta_{Ns}] = \gamma_T - \gamma_N = b;$$

$$MSE(\hat{\delta}_{Ts}) = E\left[(\hat{\delta}_{Ts} - \delta_{Ns})^2\right] = b^2 + \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.$$

$$bias(\hat{\delta}_{Ns}) = E[\hat{\delta}_{Ns} - \delta_{Ts}] = \gamma_N - \gamma_T = -b;$$

$$MSE(\hat{\delta}_{Ns}) = E\left[(\hat{\delta}_{Ns} - \delta_{Ts})^2\right] = b^2 + \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ns}$$

(6)

Third, if the bias of the estimator above is known, we can subtract the bias from $\hat{\delta}_{Ts}$ or $\hat{\delta}_{Ns}$ to get an unbiased estimate of the true trend in the other test (e.g., $\hat{\delta}'_{Ts} = \hat{\delta}_{Ts} - b$), with

$$bias(\hat{\delta}'_{Ts}) = E[\hat{\delta}_{Ts} - b - \delta_{Ns}] = 0;$$

$$MSE(\hat{\delta}'_{Ts}) = E\left[(\hat{\delta}_{Ts} - b - \delta_{Ns})^2\right] = \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.$$

$$bias(\hat{\delta}'_{Ns}) = E[\hat{\delta}_{Ns} + b - \delta_{Ts}] = 0;$$

$$MSE(\hat{\delta}'_{Ns}) = E\left[(\hat{\delta}_{Ns} + b - \delta_{Ts})^2\right] = \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ns}.$$

(7)

Fourth, we can construct a Bayesian estimate of δ_{Ns} from the observed $\hat{\delta}_{Ts}$ and the parameters of the joint distribution in (3) above:

$$\begin{aligned}
\delta_{Ns}^{*} &= E\left[\delta_{Ns}\middle|\hat{\delta}_{Ts}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ts}\right] \\
&= \gamma_N + \frac{\tau_{01}}{\tau_1 + \sigma_{Ts}}\left(\hat{\delta}_{Ts} - \gamma_T\right) \\
v_{Ns}^{*} &= Var\left(\delta_{Ns}\middle|\hat{\delta}_{Ts}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ts}\right) \\
&= \tau_0 - \frac{\tau_{01}^2}{\tau_1 + \sigma_{Ts}} \\
&= \tau_0(1 - \lambda_{Ts} r^2)
\end{aligned} \tag{8}$$

where $\lambda_{Ts} = \frac{\tau_1}{\tau_1 + \sigma_{Ts}}$ is the reliability of $\hat{\delta}_{Ts}$ as an estimate of δ_{Ts} and $r = \frac{\tau_{01}}{\sqrt{\tau_0 \cdot \tau_1}}$ is the correlation between δ_N and δ_T. Similarly, we get

$$\begin{aligned}
\delta_{Ts}^{*} &= E\left[\delta_{Ts}\middle|\hat{\delta}_{Ns}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ns}\right] = \gamma_T + \frac{\tau_{01}}{\tau_0 + \sigma_{Ns}}\left(\hat{\delta}_{Ns} - \gamma_N\right) \\
v_{Ts}^{*} &= Var\left[\delta_{Ts}\middle|\hat{\delta}_{Ns}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ns}\right] = \tau_1(1 - \lambda_{Ns} r^2)
\end{aligned} \tag{9}$$

These estimators will be shrunken toward the average change, with biases given by[6]

$$\begin{aligned}
bias(\delta_{Ns}^{*}|\delta_{Ns}) &= E[\delta_{Ns}^{*} - \delta_{Ns}|\delta_{Ns}] \\
&= E\left[\gamma_N + \frac{\tau_{01}}{\tau_1 + \sigma_{Ts}}\left(\hat{\delta}_{Ts} - \gamma_T\right) - \delta_{Ns}\middle|\delta_{Ns}\right] \\
&= (\lambda_{Ts} r^2 - 1)(\delta_{Ns} - \gamma_N) \\
bias(\delta_{Ts}^{*}|\delta_{Ts}) &= (\lambda_{Ns} r^2 - 1)(\delta_{Ts} - \gamma_T).
\end{aligned} \tag{10}$$

[6] Note that the bias of the Empirical Bayes estimates will be zero on average (assuming that the correlation between the reliability of one test and the true change in the other test is 0), but will vary across states. The variance of the bias in δ_{Ns}^{*} will be approximately $\left[\left(\bar{\lambda}_T r^2 - 1\right)^2 + r^4 var(\lambda_{Ts})\right]\tau_0$; for δ_{Ts}^{*}, it will be approximately $\left[\left(\bar{\lambda}_N r^2 - 1\right)^2 + r^4 var(\lambda_{Ns})\right]\tau_1$.

The MSE of these estimators is given by their posterior variances, v_{Ns}^* and v_{Ts}^*, respectively (Equations 8 and 9 above).

Fifth, we can combine the information in both $\hat{\delta}_{Ns}$ and $\hat{\delta}_{Ts}$ to provide Empirical Bayes estimates of both δ_{Ns} and δ_{Ts}. Assuming $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}_s$ are multivariate normal, Bayes theorem gives us

$$\begin{aligned}\boldsymbol{\delta}_s^* &= E\left[\boldsymbol{\delta}_s \middle| \widehat{\boldsymbol{\delta}}_s, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{\sigma}_s\right] \\ &= \boldsymbol{\Gamma} + \boldsymbol{\tau}(\boldsymbol{\tau} + \boldsymbol{\sigma}_s)^{-1}\left(\widehat{\boldsymbol{\delta}}_s - \boldsymbol{\Gamma}\right) \\ &= \boldsymbol{\Gamma} + \boldsymbol{\Lambda}_s\left(\widehat{\boldsymbol{\delta}}_s - \boldsymbol{\Gamma}\right) \\ \mathbf{v}_s^* &= Var\left[\boldsymbol{\delta}_s \middle| \widehat{\boldsymbol{\delta}}_s, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{\sigma}_s\right] \\ &= \boldsymbol{\Lambda}_s \boldsymbol{\sigma}_s,\end{aligned} \tag{11}$$

where $\boldsymbol{\Lambda}_s = \boldsymbol{\tau}(\boldsymbol{\tau} + \boldsymbol{\sigma}_s)^{-1}$ is the multivariate reliability matrix of $\widehat{\boldsymbol{\delta}}_s$. We can show that the diagonal elements of $\mathbf{v}_s^*$ are

$$\begin{aligned}v_{0s}^* &= Var\left(\delta_{Ns} \middle| \widehat{\boldsymbol{\delta}}_s, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{\sigma}_s\right) = \tau_0(1 - \lambda_{Ts} r^2)\left(\frac{1 - \lambda_{Ns}}{1 - \lambda_{Ns}\lambda_{Ts} r^2}\right) \\ v_{1s}^* &= Var\left(\delta_{Ts} \middle| \widehat{\boldsymbol{\delta}}_s, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{\sigma}_s\right) = \tau_1(1 - \lambda_{Ns} r^2)\left(\frac{1 - \lambda_{Ts}}{1 - \lambda_{Ns}\lambda_{Ts} r^2}\right).\end{aligned} \tag{12}$$

The Empirical Bayes estimator $\boldsymbol{\delta}_s^*$ will have bias[7]

$$\begin{aligned}bias(\boldsymbol{\delta}_s^* | \boldsymbol{\delta}_s) &= E[\boldsymbol{\delta}_s^* - \boldsymbol{\delta}_s | \boldsymbol{\delta}_s] \\ &= E\left[\boldsymbol{\Gamma} + \boldsymbol{\Lambda}_s\left(\widehat{\boldsymbol{\delta}}_s - \boldsymbol{\Gamma}\right) - \boldsymbol{\delta}_s\right] \\ &= (\boldsymbol{\Lambda}_s - \mathbf{I})\mathbf{u}_s.\end{aligned} \tag{13}$$

It is useful to compare the MSE of the estimators:

[7] Note that the bias of $\boldsymbol{\delta}^*$ will be zero on average (assuming that $Cov(\boldsymbol{\Lambda}_s, \mathbf{u}_s) = \mathbf{0}$), but will vary across states. The variance of the bias in $\boldsymbol{\delta}_s^*$ will be given by the diagonal elements of $(\overline{\boldsymbol{\Lambda}} - \mathbf{I})\boldsymbol{\tau}(\overline{\boldsymbol{\Lambda}} - \mathbf{I})' + E(\boldsymbol{\Lambda}_s - \overline{\boldsymbol{\Lambda}})\boldsymbol{\tau}(\boldsymbol{\Lambda}_s - \overline{\boldsymbol{\Lambda}})'$.

$$MSE(\hat{\delta}_{Ns}) = E\left[(\hat{\delta}_{Ns} - \delta_{Ns})^2\right] = \sigma_{Ns}$$

$$MSE(\hat{\delta}_{Ts}) = E\left[(\hat{\delta}_{Ts} - \delta_{Ns})^2\right] = (\gamma_T - \gamma_N)^2 + \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.$$

$$MSE(\hat{\delta}'_{Ts}) = E\left[(\hat{\delta}_{Ts} - (\gamma_T - \gamma_N) - \delta_{Ns})^2\right] = \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.$$

$$MSE(\delta^*_{Ns}) = \tau_0(1 - \lambda_{Ts}r^2)$$

$$MSE(\delta^*_{0s}) = \tau_0(1 - \lambda_{Ts}r^2)\left(\frac{1 - \lambda_{Ns}}{1 - \lambda_{Ns}\lambda_{Ts}r^2}\right)$$

(14)

From (14), it is straightforward to show that

$$MSE(\hat{\delta}_{Ts}) \geq MSE(\hat{\delta}'_{Ts}) \geq MSE(\delta^*_{Ns}) \geq MSE(\delta^*_{0s})$$

and

$$MSE(\hat{\delta}_{Ns}) \geq MSE(\delta^*_{0s}).$$

(15)

That is, the more information we use to estimate δ_{Ns}, the smaller the MSE. The estimator $\hat{\delta}_{Ts}$ uses no additional information; $\hat{\delta}'_{Ts}$ relies on $\hat{\delta}_{Ts}$ and $\mathbf{\Gamma}$; $\hat{\delta}^*_{Ns}$ relies on $\hat{\delta}_{Ts}$, $\mathbf{\Gamma}$, $\boldsymbol{\tau}$, and σ_{Ts}; $\hat{\delta}^*_{0s}$ relies on $\widehat{\boldsymbol{\delta}}_s$, $\mathbf{\Gamma}$, $\boldsymbol{\tau}$, and $\boldsymbol{\sigma}_s$. In each case, using additional prior information sharpens our estimate of δ_{Ns}.[8]

[8] The middle inequality is proven here:

$$\begin{aligned} MSE(\hat{\delta}'_{Ts}) - MSE(\delta^*_{Ns}) &= \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts} - \tau_0(1 - \lambda_{Ts}r^2) \\ &= \tau_1 - 2\tau_{01} + \sigma_{Ts} - \tau_0\lambda_{Ts}r^2 \\ &= \tau_1 - 2r\sqrt{\tau_0\tau_1} + \tfrac{1-\lambda_{Ts}}{\lambda_{Ts}}\tau_1 - \tau_0\lambda_{Ts}r^2 \\ &= \tau_0\left(\frac{\tau_1}{\lambda_{Ts}\tau_0} - 2\sqrt{\frac{\tau_1}{\tau_0}}r - \lambda_{Ts}r^2\right) \\ &= \tau_0\left(\sqrt{\frac{\tau_1}{\lambda_{Ts}\tau_0}} - \sqrt{\lambda_{Ts}}r\right)^2 \\ &\geq 0. \end{aligned}$$

Where the equality holds *iff*

The inequalities in (15) show that, although the Bayesian estimates are biased, they have smaller MSE than any of the other estimators. The optimal *unbiased* estimator of δ_{Ns} will be $\hat{\delta}_{Ns}$, though it will have larger MSE than δ^*_{0s}, the Bayesian estimate based on both the NAEP and state test observed trends. Once we know $\mathbf{\Gamma}$ and $\boldsymbol{\tau}$ (which we obtain by fitting model (4) above), we can compare the implied bias and MSE of the estimators.

Primary Results

In order to compute biases, correlations, and root mean squared errors (RMSEs) implied by the modeling framework, we fit the model in Equation 4 to NAEP and state-test 2-year changes in math and reading for grades 4 and 8. We assume no linking error in our initial specification. We compare two analytic samples, one consisting of all year-pairs from 2009-2024 (but excluding 2019-2022, which spans three years and reflects unusually large pandemic-era changes), and another consisting of year-pairs from 2015-2024 (again excluding 2019-2022). This yields 628 paired observations of $\hat{\delta}_{Ns}$and $\hat{\delta}_{Ts}$ for the longer panel and 327 for the shorter panel, respectively. Both samples also include additional year-pairs that contribute to a NAEP change with no state counterpart, 596 in the longer panel and 285 in the shorter panel. Table 3 reports the estimated variance components $(\tau_0, \tau_1, \tau_{01})$, reliabilities, mean differences ("bias"), and the implied RMSE of several estimators of the true NAEP change, δ_{Ns}.

We briefly discuss here eight substantively important patterns in Tables 3 and 4. First, the estimated variance of true state-test changes is consistently larger than the variance of true NAEP changes ($\tau_1 > \tau_0$). This may partly reflect linking error variance in the state test changes that the model does not account for, but the estimated difference between $\tau_1$and τ_0 is too large to be entirely due to unmeasured

$$r = \frac{1}{\lambda_{Ts}}\sqrt{\frac{\tau_1}{\tau_0}}$$

error variance in state test trends, given the estimates of τ_0 and τ_{01}.[9] The additional variance of state test changes may reflect construct-relevant heterogeneity (states differ in how much they improve on the knowledge and skills emphasized by their state assessments, beyond what NAEP trends measure) and/or construct-irrelevant heterogeneity (states differ in the extent of score inflation mechanisms, e.g., differential motivation, coaching, exclusions, or misconduct, affecting state tests more variably than NAEP).

Second, under the assumption that there is no state linking error, the estimated true correlation between NAEP and state-test changes is moderate and far from unity. In the full specification, the correlation is 0.45, and in recent years, the correlation is 0.58 (it is higher in 2017-2019 and 2022-2024). This indicates that, even when both testing programs are stable, they are not measuring the same latent changes. We show, however, in the next section that a more realistic estimate of state test linking error variance raises the estimated correlation substantially.

Third, the reliability of NAEP changes is low. In pooled specifications, NAEP reliabilities are 0.48 overall and 0.45 in recent years[10]. The reliability of state test score changes will, of course, depend on the extent of linking error. Under the assumption that state test score trends have no linking error, state-test change reliability is near unity (≈0.99). This difference is explained largely by NAEP's sampling design (state-level samples) versus near-census state testing and, in small part, by the larger true variance of state-test changes. Again, however, our analyses in the next section show that a more realistic estimate of the degree of linking error lowers the implied reliability of state test trends substantially.

---

[9] To see this, note that the ratio $\frac{\tau_0}{\tau_1} = \left(\frac{\tau_0}{\tau_{01}} r\right)^2$, where r is the correlation between the true NAEP and state tests. Because $|r| \leq 1$, $\frac{\tau_0}{\tau_1}$ has a maximum value of $\left(\frac{\tau_0}{\tau_{01}}\right)^2$. Given that $\tau_0 < \tau_{01}$ in most columns of Tables 3 and 4, the ratio $\frac{\tau_0}{\tau_1}$ will be less than 1, implying that τ_1 (the variance of true state test score changes) is larger than τ_0 (the variance of true NAEP test score changes) regardless of the correlation between the true test score changes.

[10] We report the average of state reliabilities for interpretability. An alternative approach would be to report the reliability for the average state's error variance. The difference is small for NAEP (0.43 vs. 0.45 for the 2015-2024 pooled model), and negligible for state reliabilities (0.9873 vs. 0.9874 for the 2015-2024 pooled model).

Fourth, state tests tend to report more positive changes than NAEP. In the pooled specification, the average state-NAEP discrepancy ("bias") is 0.051 SD units overall and 0.047 SD units in recent years. This systematic bias is consistent with both desirable alignment to state tests over NAEP and test-score inflation theories. We treat this here as a descriptive parameter rather than the result of specific causal mechanisms.

Table 3. Model parameter estimates for state and NAEP trends, 2009-2024, pooled by year models, assuming no linking error.

	Pooled (2015-19, 2022-24)	Pooled (2009-19, 2022-24)	2009-11	2011-13	2013-15	2015-17	2017-19	2019-22	2022-24
Tau Matrix									
Var(NAEP change)	0.00164	0.00173	0.00150	0.00182	0.00212	0.00157	0.00194	0.00308	0.00130
Var(State change)	0.00789	0.00911	0.01046	0.01069	0.00967	0.00521	0.00993	0.01080	0.00866
Cov(NAEP, State)	0.00210	0.00179	0.00175	0.00146	0.00099	0.00110	0.00327	0.00383	0.00197
True Correlation	0.58	0.45	0.44	0.33	0.22	0.39	0.74	0.66	0.59
Reliabilities									
NAEP	0.45	0.48	0.46	0.54	0.50	0.45	0.53	0.60	0.35
State	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
Average Change									
NAEP	-0.018	-0.005	0.026	0.033	-0.033	-0.013	-0.035	-0.143	-0.006
State	0.029	0.046	0.094	0.043	0.053	0.023	0.019	-0.162	0.044
Bias									
NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State	0.047	0.051	0.068	0.009	0.086	0.036	0.055	-0.019	0.049
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State & NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SD of bias									
State*	0.03687	0.03991	0.03742	0.04205	0.04581	0.03835	0.03604	0.04692	0.03325
State & NAEP	0.02154	0.02152	0.02070	0.02007	0.02342	0.02168	0.01952	0.02290	0.02181
Correlation with True NAEP Change									
NAEP	0.67	0.69	0.68	0.72	0.73	0.66	0.72	0.77	0.59
State	0.58	0.45	0.44	0.33	0.22	0.38	0.74	0.66	0.58
State - Bias	0.58	0.45	0.44	0.33	0.22	0.38	0.74	0.66	0.58
State*	0.58	0.45	0.44	0.33	0.22	0.38	0.74	0.66	0.58
State & NAEP	0.75	0.73	0.72	0.75	0.71	0.70	0.84	0.83	0.72
RMSE									
NAEP	0.047	0.045	0.043	0.042	0.045	0.046	0.043	0.047	0.051
State	0.087	0.100	0.115	0.099	0.131	0.077	0.092	0.077	0.093
State - Bias	0.074	0.086	0.093	0.099	0.099	0.068	0.074	0.080	0.078
State*	0.033	0.037	0.035	0.040	0.045	0.037	0.030	0.042	0.029
State & NAEP	0.027	0.028	0.027	0.028	0.032	0.028	0.024	0.031	0.025
N obs	939	1,852	337	346	230	309	297	326	333
Npairs	327	628	133	142	26	105	93	122	129

Note: This table assumes no linking error. All models pool over subjects (math and reading) and grades (4 and 8). The pooled models include subject x grade x year fixed effects, and the single year-pair models include subject x grade fixed effects.

Fifth, the relationships among bias, correlation, and precision clarify why state trends can be simultaneously "biased" yet also informative about true NAEP change (as measured by their correlation with the true NAEP change). Despite the positive mean discrepancy, the observed state-test changes remain correlated with true NAEP changes because (i) the true correlation r_{NT} is moderate and (ii) state-test changes are measured with high precision. By contrast, NAEP changes are unbiased but comparatively noisy, which limits how informative they can be about true NAEP changes.

Sixth, the RMSE comparisons in Table 3 quantify the consequences for estimation of δ_{Ns}. Following Equation 14, the naive estimator $\hat{\delta}_{Ts}$ (using raw state changes as NAEP changes) has a large RMSE because it combines squared bias, the variance of true trend discrepancies, and the measurement error of the state trend. Bias-correcting state changes removes the squared bias component and meaningfully reduces RMSE, but the bias-corrected state estimator remains substantially less accurate than NAEP because the variance of true trend discrepancies remains. In contrast, in the pooled specifications and in most year-pairs, the Empirical Bayes (shrinkage) estimator based on state information alone has smaller RMSE than either the NAEP-only estimator or the bias-corrected state-only estimator. Intuitively, because $r_{NT} < 1$, even true state changes contain a substantial test-specific component that does not predict true NAEP change. The EB estimator effectively regresses state changes toward the NAEP mean by the factor $\tau_{01}/(\tau_1 + \sigma_T)$, exploiting the state test's high precision while discounting test-specific variation.

Seventh, Table 4 shows results separately by subject and grade, pooled over both 2009-2024 and 2015-2024, excluding the 3-year 2019-2022 pandemic period. Average trend discrepancies are similar across subjects and grades. However, true correlations between state and NAEP trends are generally lower for Reading Language Arts than for mathematics. This raises questions about whether distinctions between NAEP Reading and state tests of English Language Arts, which sometimes incorporate information about progress in other domains, might cause disagreements in trends.

Table 4. Model parameter estimates by subject and grade, pooled over years, assuming no linking error.

	2015-19, 2022-24				2009-19, 2022-24			
	Math 4	RLA 4	Math 8	RLA 8	Math 4	RLA 4	Math 8	RLA 8
Tau Matrix								
Var(NAEP change)	0.00266	0.00144	0.00129	0.00129	0.00257	0.00162	0.00163	0.00119
Var(State change)	0.00603	0.00992	0.00582	0.00926	0.00788	0.01092	0.00652	0.01053
Cov(NAEP, State)	0.00269	0.00254	0.00177	0.00121	0.00275	0.00139	0.00190	0.00117
True Correlation	0.67	0.67	0.65	0.35	0.61	0.33	0.58	0.33
Reliabilities								
NAEP	0.54	0.40	0.42	0.41	0.55	0.45	0.51	0.40
State	0.98	0.99	0.98	0.99	0.99	0.99	0.98	0.99
Average Change								
NAEP	0.018	-0.036	-0.013	-0.041	0.013	-0.009	-0.009	-0.013
State	0.044	0.025	0.028	0.019	0.053	0.046	0.037	0.048
Bias								
NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State	0.026	0.062	0.041	0.060	0.040	0.055	0.046	0.061
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State & NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SD of bias								
State*	0.04138	0.03365	0.03145	0.03541	0.04420	0.03967	0.03643	0.03405
State & NAEP	0.02332	0.02062	0.01921	0.02114	0.02278	0.02219	0.01966	0.02050
Correlation with True NAEP Change								
NAEP	0.73	0.63	0.64	0.63	0.74	0.67	0.69	0.63
State	0.67	0.67	0.64	0.35	0.61	0.33	0.58	0.33
State - Bias	0.67	0.67	0.64	0.35	0.61	0.33	0.58	0.33
State*	0.67	0.67	0.64	0.35	0.61	0.33	0.58	0.33
State & NAEP	0.82	0.77	0.77	0.67	0.80	0.69	0.78	0.66
RMSE								
NAEP	0.049	0.048	0.045	0.045	0.047	0.046	0.043	0.044
State	0.064	0.101	0.073	0.109	0.082	0.113	0.081	0.115
State - Bias	0.058	0.080	0.061	0.091	0.071	0.099	0.067	0.097
State*	0.038	0.028	0.028	0.034	0.040	0.038	0.033	0.033
State & NAEP	0.030	0.024	0.023	0.027	0.030	0.029	0.025	0.026
N obs	245	244	216	234	475	476	433	468
Npairs	92	91	63	81	169	170	127	162

Note: This table assumes no linking error. All models pool over the year-pairs in the indicated range, excluding 2019-22, and include year fixed effects.

Eighth and finally, the modeling framework also yields a fifth estimator: the Empirical Bayes estimator that conditions on *both* observed signals, $E[\delta_{Ns} \mid \hat{\delta}_{Ns}, \hat{\delta}_{Ts}]$. This two-signal EB estimator is the

posterior mean under the multivariate normal model and has the smallest MSE among the estimators considered, because it uses strictly more information than the state-only shrinkage estimator. In the pooled specification, adding NAEP information on top of state information produces an additional reduction in RMSE, as expected based on Equation (15) above. This estimator is particularly relevant for retrospective estimation of historical NAEP trends in windows where both NAEP and state assessments are available. In contrast, the state-only shrinkage estimator is most relevant for both retrospective and prospective prediction of 1-year NAEP-scale changes in windows where NAEP is unavailable but state testing continues within a stable regime, under the assumption that sequential 1-year changes are independent.[11] We test the viability of this in practice in a later section.

Taken together, these results support three conclusions. First, relying on unadjusted state-test changes as if they were NAEP changes yields estimates with substantial bias and high RMSE. Second, bias correction helps by recentering state trends to match NAEP on average, but substantial disagreement remains because the two tests' true trends differ across states and years, not just on average. Third, Empirical Bayes shrinkage estimators, especially the two-signal EB estimator when both "watches" are available, can deliver lower-RMSE estimates of true NAEP change than NAEP changes alone, despite the presence of systematic state-test bias.

Accounting for Linking Error in State Test Trends

In Tables 3 and 4 above, we assumed no linking error (neither equating error nor discretization error) in state tests. In this section, we assess the plausibility of this assumption and the sensitivity of our estimates to the presence of equating error. We focus here on equating error, as we show in Appendix A

[11] Because NAEP is available in all states (every other year), this may seem like a case that does not occur. But of course, in non-NAEP years, we may observe changes on a stable state test despite the absence of NAEP change estimates. In this case (under some assumptions about the independence of sequential 1-year changes), the EB estimator that uses observed state changes to estimate NAEP changes can be used.

that discretization error does not make a substantial contribution to linking error variance, at least in a limited set of states for which we can estimate it. Estimates of the magnitude of equating error are not readily available. Although the *Standards for Educational and Psychological Testing* recommend that "standard errors of equating functions should be estimated and reported whenever possible" (American Educational Research Association et al., 2014, p. 105), we were not able to find such estimates in technical reports for state testing programs. One available estimate comes from Michaelides and Haertel (2014), who use a bootstrap approach to estimate a standard error of equating of 0.029 on a scale with a 0.664 standard deviation, implying an equating error variance of approximately 0.002 in squared standard deviation units per equating operation for a state test with 44 common items embedded across 8 matrix-sampled forms. Because a 2-year change involves two such equating operations, and assuming equating errors across adjacent years are approximately uncorrelated, a plausible approximation of equating error variance for a test like this is 0.004 per 2-year change. However, the Michaelides and Haertel (2014) estimate is based on data from a single state and a single year's equating, so may not be generalizable.

Given the lack of information on the magnitude of equating error, we conduct a sensitivity analysis in which we add a range of values of linking error variance to the state test change measurement-error variances before fitting the model. We then compare the estimates from these models to those shown in Table 3 (column 1), where we assumed no linking error variance in state tests. Following this, we develop and apply a method for using both NAEP and state test data to gauge the magnitude of equating error in state assessments. We then use these estimates of equating error variance to infer more appropriate estimates of the parameters shown in Tables 3 and 4.

Table 5 presents a sensitivity analysis in which we add fixed amounts of error variance (ranging from 0.001 to 0.005) to the state test change measurement-error variances before fitting the model. The results show that state test change reliability falls substantially as assumed linking error increases, from 0.99 in the base case to 0.36 under the most extreme assumption. The estimated true correlation rises

correspondingly (from 0.58 to 0.96), reflecting disattenuation. Critically, however, the estimated bias, RMSE, and shrinkage estimates themselves are unchanged across all specifications. Each of the five estimators of the true NAEP change and its MSE depends on τ_1 and σ_{Ts} only through their sum, the total

Table 5. Sensitivity of model estimates to underestimated state test score trend imprecision.

	Original	1	2	3	4	5
Added Variance	0.000	0.001	0.002	0.003	0.004	0.005
Tau Matrix						
Var(NAEP change)	0.00164	0.00164	0.00164	0.00164	0.00164	0.00164
Var(State change)	0.00789	0.00689	0.00589	0.00489	0.00388	0.00289
Cov(NAEP, State)	0.00210	0.00210	0.00209	0.00209	0.00209	0.00209
True Correlation	0.58	0.62	0.67	0.74	0.83	0.96
Reliabilities						
NAEP	0.45	0.45	0.45	0.45	0.45	0.45
State	0.99	0.86	0.74	0.61	0.49	0.36
Average Change						
NAEP	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018
State	0.029	0.029	0.029	0.029	0.029	0.029
Bias						
NAEP	0.000	0.000	0.000	0.000	0.000	0.000
State	0.047	0.047	0.047	0.047	0.047	0.047
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000
State*	0.000	0.000	0.000	0.000	0.000	0.000
State & NAEP	0.000	0.000	0.000	0.000	0.000	0.000
SD of bias						
State*	0.03687	0.03634	0.03565	0.03469	0.03323	0.03067
State & NAEP	0.02154	0.02084	0.02012	0.01939	0.01863	0.01780
Correlation with True NAEP Change						
NAEP	0.67	0.67	0.67	0.67	0.67	0.67
State	0.58	0.58	0.58	0.58	0.58	0.58
State - Bias	0.58	0.58	0.58	0.58	0.58	0.58
State*	0.58	0.58	0.58	0.58	0.58	0.58
State & NAEP	0.75	0.75	0.75	0.75	0.75	0.75
RMSE						
NAEP	0.047	0.047	0.047	0.047	0.047	0.047
State	0.087	0.087	0.087	0.087	0.087	0.087
State - Bias	0.074	0.074	0.074	0.074	0.074	0.074
State*	0.033	0.033	0.033	0.033	0.033	0.033
State & NAEP	0.027	0.027	0.027	0.027	0.027	0.027
N obs	939	939	939	939	939	939
Npairs	327	327	327	327	327	327

Note: The specification/sample in this table is identical to Column 1 of Table 3.

observed variance of state test score trends (see Equations (8), (11), and (14)). Assuming different values for linking error variance repartitions this sum but does not change it. Therefore, although the state test trend reliability will be lower and the state-NAEP true correlation may be higher to the extent that such errors are present and unmodeled, estimates of the true NAEP trend, and their RMSEs, are robust to plausible magnitudes of measurement error in state test trend estimation. Nonetheless, the reliability of state test score trends and the implied correlation between state test score trends and NAEP trends are important quantities to estimate.

We next estimate the magnitude of linking error in state assessments, using both state test and NAEP trend data. Conceptually, we do this as follows: we first estimate the variance of state test score changes across grades within a state-year-subject. This variance includes a) true variance across grades; b) linking error variance; and c) sampling error variance. We have an estimate of the sampling error variance in state tests from the HETOP estimation; given the large sample sizes, this error is very small. We can estimate the linking error variance by subtracting the true cross-grade variance of changes—if we know it—from the observed variance. We obtain upper and lower bounds of the true cross-grade variance using NAEP data and subtract these from the observed cross-grade variance.

We first standardize 2-year test score changes (for both NAEP and state tests), so they both are measured in state-specific standard deviation units. Using the NAEP changes for each 2-year period (denoted $\hat{\delta}_{sbg}$ for the change in state s, subject b and grade g), we fit the following two precision-weighted models, nesting the 4 NAEP subject-grades in states:

$$\begin{aligned}
\hat{\delta}_{sbg} &= \alpha_{sbg} + \epsilon_{sbg} \\
\alpha_{sbg} &= \beta_{1s}(M) + \beta_{2s}(R) + r_{sbg} \\
\beta_{1s} &= \gamma_{10} + u_{1s} \\
\beta_{2s} &= \gamma_{20} + u_{2s} \\
\epsilon_{sbg} &\sim \left(0, v_{sbg}\right); r_{sbg} \sim (0, \sigma^2); \mathbf{u}_s \sim (\mathbf{0}, \boldsymbol{\tau}^2)
\end{aligned} \tag{16}$$

and

$$
\begin{aligned}
\hat{\delta}_{sbg} &= \alpha_{sbg} + \epsilon_{sbg} \\
\alpha_{sbg} &= \beta_{1s}(M) + \beta_{2s}(R) + \beta_{3s}(M8) + \beta_{4s}(R8) + r_{sbg} \\
\beta_{1s} &= \gamma_{10} + u_{1s} \\
\beta_{2s} &= \gamma_{20} + u_{2s} \\
\beta_{3s} &= \gamma_{30} \\
\beta_{4s} &= \gamma_{40} \\
\epsilon_{sbg} &\sim \left(0, v_{sbg}\right); r_{sbg} \sim (0, \sigma^2); \mathbf{u}_s \sim (\mathbf{0}, \boldsymbol{\tau}^2)
\end{aligned}
\tag{17}
$$

Here M and R are dummies indicating math or reading; $M8$ and $R8$ are dummies for math and reading 8th grade; and v_{sbg} is the known sampling error variance of $\hat{\delta}_{sbg}$.

The parameter of interest in these models is σ^2. The σ^2 from model (16) is between-grade variance in 2-year NAEP score changes, within state-subjects. It includes linking error variance due to differences in linking error between 4th and 8th grade and true between grade variance in changes. The σ^2 from model (17) is between-grade variance in 2-year NAEP score changes, net of common national grade-specific trends, within state-year-subjects. It excludes linking error variance due to systematic differences in linking error between 4th and 8th grade, but it also excludes true variance between grades that is common across states within a year-subject. So the σ^2's from models (16) and (17) represent upper and lower bounds on the true within-state-subject trend variance, respectively. We fit this model for each pair of NAEP years, from 2003 to 2024 (excluding 2019-2022), and then average the σ^2's across years.

We then fit model (17) using state test score changes for 2-year periods from 2009-2024 (excluding 2019-2022). When using state test score changes, the σ^2 reflects both true-within state-subject variance in trends and linking error variance, excluding variance in trends that is common across states within grade-subjects. So σ^2 here is an upper bound on the 2-year linking error variance of state trends.

Table 6 reports the variance estimates from these models. Subtracting the variance estimates from the NAEP models from the Model 17 State (all grades) estimates, we get lower and upper bounds of linking error variance of 0.00308 and 0.00382. We use the estimate based on all grades for the state test

trend variance, because it relies on more information, but the estimated variance is virtually identical if we use just 4th and 8th grade in the model. Given these estimates, we use 0.003 and 0.004 as plausible lower- and upper-bounds of state two-year linking error variance. These estimates are consistent with the Michaelides and Haertel (2014) estimates, bolstering our confidence in them.

Table 6. Estimates of between-grade variance from 2-year NAEP and state tests.

Model	NAEP	State (all grades)	State (grades 4 & 8 only)
16	0.00111		
17	0.00037	0.00419	0.00420

As Table 5 shows, these estimates imply that the reliability of state test trends is 0.49 to 0.61, modestly higher than the NAEP reliability of 0.45, but far below what we estimate when ignoring linking error. They also imply the correlation between true state and NAEP trends is 0.74 to 0.83, much higher than we estimate when ignoring linking error.

Validation of the EB estimator using out-of-sample predictions

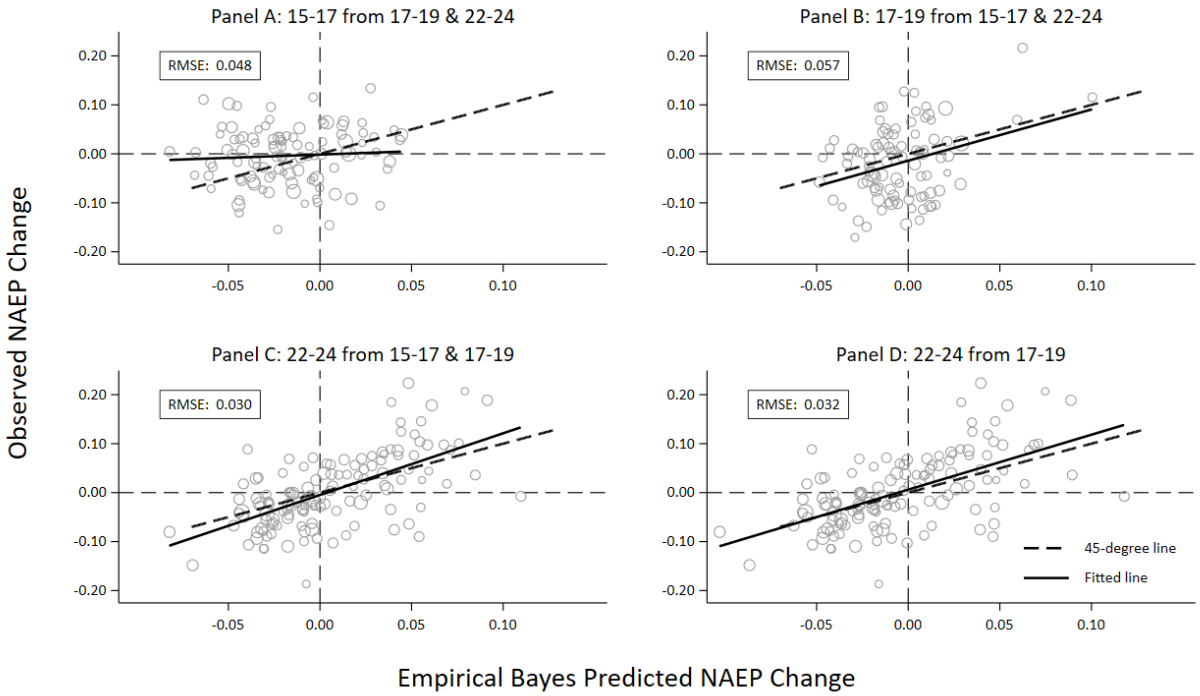
The EB estimators derived above require estimates of the model parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\tau}$. These must be estimated from other years, grades, or states when no observed trend is available for the target window. Whether estimating NAEP trends in non-NAEP years from stable state test trends, or when estimating state trends across a test redesign, the EB approach depends on whether $\boldsymbol{\Gamma}$ and $\boldsymbol{\tau}$ are sufficiently stable across contexts to support accurate out-of-sample prediction. We evaluate this by testing how well the state-only shrinkage estimator predicts NAEP trends in year-pairs whose data were not used to fit the model.

We conduct three leave-one-out prediction exercises using the three year-pairs in our primary analytic sample (2015-17, 2017-19, and 2022-24). In each exercise, we fit the model on two of the three year-pairs, compute the state-only EB shrinkage estimate for the held-out year-pair using the fitted model parameters, and compare those predictions to the observed NAEP changes in the held-out year-pair. We

also conduct a fourth "forward prediction" exercise in which we fit the model on 2017-19 data only and predict 2022-24 NAEP changes, mimicking a situation in which only one prior year-pair of data is available. If the model parameters ($\boldsymbol{\tau}$ and $\boldsymbol{\Gamma}$) are invariant across years, the out-of-sample EB shrinkage estimate is an unbiased linear predictor of the true NAEP change, implying that regressing the observed NAEP change on the EB estimated true NAEP change should yield the identity line $y = x$.

The out-of-sample predictions are reported in Figure 1. Forward predictions of 2022-24 NAEP changes perform best, with RMSEs of 0.030 in Panel C (trained on 2015-17 and 2017-19) and 0.032 in Panel D (trained on 2017-19 only). The near identical RMSEs across the two panels suggest that little is gained by adding the earlier 2015-2017 period to the training data. Moreover, the RMSEs of 0.030 and 0.032 are consistent with the RMSEs of 0.030 and 0.037 from the training models that use only 2017-19 or 2015-17 data, respectively (shown in Table 3), suggesting that state-only shrinkage estimators can match or exceed the precision of NAEP alone in held-out years. Predicting 2015-17 and 2017-19 NAEP

Figure 1. Observed NAEP Changes Versus Out-of-Sample Predicted NAEP Changes



changes from other years is less successful (RMSEs are 0.048 and 0.057, respectively), though even in these cases, the RMSE of the prediction is only slightly larger than the RMSE we would obtain if we used the actual NAEP estimate in these cases (where the RMSEs are 0.046 and 0.043, respectively, in Table 3). Taken together, these results demonstrate that the state-only shrinkage estimator can predict true NAEP changes with moderate accuracy in held-out years, providing empirical support for its use in years when NAEP is unavailable.

Discussion and Conclusion

This paper contributes empirical findings and methodological tools to improve educational progress monitoring using multiple tests. We estimate the relationship between state test score trends and NAEP trends from 2009-2024. We find that two-year NAEP and state test trends are both only moderately reliable (reliability is 0.45 for NAEP; 0.49-0.61 for state tests). The two trends are correlated at roughly 0.80, based on our estimates of state test linking error variance. State test trends are more positive, on average, than NAEP trends by approximately 0.05 SD units per two-year period, roughly half the bias documented in earlier work from selected states in the 1990s and 2000s (Ho, 2007; Jacob, 2007).

We also describe the properties of five different estimates of true NAEP trends that reveal practical applications of the modeling framework. First, in non-NAEP years, the state-only EB shrinkage estimator can predict NAEP-scale trend estimates with accuracy that matches or exceeds NAEP itself in held-out validation exercises. Second, when a state test trend breaks, NAEP trends can provide an estimate of the implied change on the state test scale. Third, when both signals are available simultaneously, the two-signal EB posterior mean strictly dominates any single-signal estimator in MSE and in correlation with the true change, improving retrospective estimates of both NAEP and state trends. Fourth, the framework generalizes beyond our NAEP-state example to any setting where multiple assessments (for example, district-administered tests or international assessments) overlap in time and

population, enabling synthesis of trend information from a richer portfolio of measures than any single program provides.

Our paper also motivates several questions that extensions of our modeling framework can address. What causes or predicts the magnitude and direction of discrepancies between state and NAEP trends? Candidate factors include differential changes in stakes, content, and administration conditions. How do trend reliability and relationships differ by subgroup? And how could we incorporate trend information from a third or fourth testing program?

Finally, our analysis reveals a haphazard and patchwork data infrastructure for educational monitoring in the United States. State test trends break often; equating error is undocumented; and dueling trends, when available, are reported separately, not synthesized. Our modeling framework is a first step. Future research, policy, and practice can improve both the estimation and the magnitude of trend reliability to improve our understanding of educational progress in the United States.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping state proficiency standards onto NAEP scales: 2005-2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pubs/studies/2010456.aspx

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313-338). Springer. https://doi.org/10.1007/978-0-387-49771-6_17.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x

Briggs, D.C. (2024). The Past, Present, and Future of Large-Scale Assessment Consortia. *Educational Measurement: Issues and Practice, 43(4)*, 62-72. https://doi.org/10.1111/emip.12634

Education Data Center. (2025). *State assessment data repository (Version 3.0): Data documentation*. Retrieved from https://www.zelma.ai/data_documentation/EDC_technical_documentation_v3.0.pdf

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher, 36*(5), 268–278. https://doi.org/10.3102/0013189X07306556

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice, 26*(4), 11–20. https://doi.org/10.1111/j.1745-3992.2007.00104.x

Ho, A. D., & Haertel, E. H. (2007). *(Over)-interpreting mappings of state performance standards onto the NAEP scale*. Council of Chief State School Officers. https://andrewho.scholars.harvard.edu/sites/g/files/omnuum4106/files/andrewho/files/ho_haertel_overinterpreting_mappings.pdf

Ho, A. D., & Polikoff, M. S. (2025). Test-based accountability in K-12 education. In M. Pitoniak and L. Cook (Eds.), *Educational Measurement* (5th ed.). Routledge.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489-517.

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Jacob, B. A. (2007). Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments (Working Paper No. 12817). National Bureau of Economic Research. https://doi.org/10.3386/w12817

Ji, C. S., Rahman, T., & Yee, D. S. (2021). *Mapping state proficiency standards onto the NAEP scales: Results from the 2019 NAEP reading and mathematics assessments* (NCES 2021-036). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED612877.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer New York.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education/Praeger.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Center for the Study of Evaluation, University of California, Los Angeles.

Mazzeo, J., Liu, B., Donoghue, J., & Xu, X. (2025). Approximate standard errors for NAEP results that incorporate linking error under the random group design. In *Innovative Digital-Based International Large-Scale Assessments: Foundations, Methodologies, and Quality Assurance* (pp. 391-417). Cham: Springer Nature Switzerland.

Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*(1), 46-57.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263–283. https://doi.org/10.1162/rest.2010.12318

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education, 118*(3), 341–368. https://doi.org/10.1086/664773

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., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics, 46*(2), 138–167.

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics, 42*(1), 3–45.

State of Georgia. (2011). Investigation into test cheating in Atlanta Public Schools. Retrieved from https://archive.org/details/215260-georgia-investigation

U.S. Department of Education. (2016). *State assessments in reading/language arts and mathematics, school year 2014-15: EDFacts data documentation*. National Center for Education Statistics. Retrieved from https://web.archive.org/web/20210926130047/https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html.

Yee, D. S., & Ho, A. D. (2015). Discreteness causes bias in percent-above-cutoff comparisons: A case study from educational testing. *The American Statistician*, *69*(2), 174-181. https://doi.org/10.1080/00031305.2015.1031828

Appendix A. Estimating discretization error in HETOP-estimated state test score trends

Yee and Ho (2015) show that shifting a proficiency cut score by one discrete score point up or down can change the estimated change in proficiency percentages substantially. Using reported state test score distributions from technical manuals in 13 states from 2010-2011, they report a root mean squared difference (RMSD) of 1.5 percentage points in math and 1.8 percentage points in Reading/English Language Arts. Using inverse normal transformations to convert this to a standardized (z) scale results in maximal discretization errors of 0.038-0.045, assuming proficiency cut scores are at the mean of a normal distribution.

To explore whether these discretization errors are a substantial percentage of the estimates we bound, we use EDC data from recent years where states report both means and standard deviations, enabling estimated differences in standard deviation units. These are available only for matched NAEP years 2022-2024, in Michigan, Minnesota, and South Carolina. We show the correlations and RMSDs in Figure A.1. Differences can arise not only because of discretization but also the (generally flexible) assumption of respective normality (Ho & Reardon, 2012). For these three states, RMSDs are generally smaller than those estimated by Yee & Ho (2015), 0.020 for math and 0.017 for reading/English language arts (0.0003-0.0004 in variance). Trends may be more robust due to our use of multiple cut scores (not just the proficient cut score) or due to this selected sample of three states. Taking estimates of variance from our bounding exercise of 0.003 to 0.004, discretization would contribute only 7-13% of the variance. While effect sizes from more states are necessary to draw stronger conclusions, we conclude that discretization error contributes only modestly to linking error and that the likely bulk of imprecision is equating error.

Figure A.1. Approximating discretization error in state test score trends (2022-2024)