Two Watches:

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

sean f. reardon
Stanford University

Andrew D. Ho
Harvard Graduate School of Education

Jie Min
Stanford University

Version May 13, 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 Discrepant 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 two watches that disagree about the time, two tests may report different trends for the same populations, subjects, and time periods. We develop a precision-adjusted multilevel measurement error model of the relationship between 1) state test score trends and 2) National Assessment of Educational Progress (NAEP) score trends for the same states. Using data from NAEP and state testing programs, the model jointly estimates the true variance in NAEP trends, the true variance in state test score trends, their respective reliabilities, their true correlation, and systematic discrepancies (bias) in state test score trends as estimates of NAEP trends.

We find that NAEP trends have low reliabilities: 0.49 for 2-year trends from 2009-2024 and 0.38 after 2015. State test trends appear to have higher reliabilities due to census testing. Across stable state testing eras, the true correlation between NAEP and state test trends is approximately 0.47 since 2009 and 0.61 since 2015. As estimates of NAEP trends, state trends show consistent upward bias of approximately 0.05 standard deviation units per two-year period. Our modeling framework accounts for both measurement error and negative serial correlation in consecutive trends. We also demonstrate how shrinkage (Empirical Bayes) estimators may produce substantially more accurate estimates of educational progress by combining information from NAEP and state assessments or, when NAEP is missing, predicting NAEP trends from state test score trends alone.

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 illuminates historical disagreements and suggests 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 state test trends exceeded NAEP trends by an average of .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).

Table 1 reviews general differences between NAEP and state tests along 10 dimensions. These are

[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).

summarized well by prior literature 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, predict 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 was limited in scope to selected years and states, and methods 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), and sampling error and other forms of error in state test score means attenuates 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.

Our modeling framework accounts for measurement error in test score trends. We have well-estimated error variances for two sources of error: sampling error in NAEP (Mislevy et al., 1992) and, for state test scores, sampling error and estimation error due to the coarsening of scores into a small number of proficiency categories (the latter applies when state test score means are estimated from proficiency category counts using a heteroskedastic ordered probit model (HETOP; Reardon et al., 2017)). Additional sources of error may also be important but are difficult to quantify precisely. Linking error—misalignment of test score scales in different test administrations—arises from the sampling of common items on which year-to-year test equating depends (Kolen & Brennan, 2014). Unlike student sampling error, linking error does not shrink with near-census testing and therefore does not vanish in state trend estimates despite near-complete population coverage. An additional source, discretization error, arises from the coarseness of test scales from which reported proficiency distributions are constructed (Ho & Yu, 2015; Yee & Ho, 2015); this can cause non-trivial errors in percent-above-cutscore estimates. When state test score means are estimated from these noisy proficiency category counts using HETOP models, the resulting estimates contain error due to the discretization.

We address two sets of research questions here. The first investigates the extent to which NAEP

and state tests measure the same underlying trends. Given the answer to this first question, we then investigate how well we can estimate true NAEP trends from either or both of the error-prone observed NAEP and state trends. Specifically, we ask:

1. How similar are NAEP and state test trends, once we account for various sources of error? What are their reliabilities, correlations, average discrepancies, and root mean squared errors?
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?

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 0.6 and 0.9, depending on the magnitude of linking and discretization 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. 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 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, improving inference about true educational progress. While biased and often broken, 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 and the EDFacts Initiative. 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. To account for NAEP initiatives to expand and standardize inclusion of English learners and students with disabilities early in this time period, we rely on the Expanded Population Estimates (EPE) of means and standard deviations provided by the National Center of Education Statistics (see Braun et al., 2010; McLaughlin, 2005; National Institute of Statistical Sciences, 2009).[2]

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 data from the Education Data Center (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.

[2] Expanded population estimates (EPEs) for NAEP 2024 were not available at the time of writing. Because EPE and regular NAEP scores differ by a small magnitude, we estimate 2024 EPEs with a mean adjustment to historical differences between EPE and regular NAEP scores, for each subject-grade combination. Note that the correlation between EPE and regular NAEP estimates are near unity (~0.99); as a result, our central substantive conclusions are unchanged if we use the regular NAEP estimates in the linking.

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 through 2015, but only a few states have availability retrospectively through 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 EDFacts surveys of state agency representatives about test changes (e.g., U.S. Department of Education, 2016).[3]

To estimate trends from state testing programs when trends are comparable, we use population-level counts of students in ordered proficiency categories from the EDFacts 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, resulting in 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 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]

[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.

[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 argue in Reardon et al. (2021), extremely consistent reliability estimates within states-subject-grade-year combinations suggests that imputation is not consequential. We do not adjust NAEP trends, as NAEP estimation procedures account for measurement error due to item assignment to examinees (Mislevy et al., 1992).

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 English Language Arts and mathematics). Comparable trends are those with no reported test score scale or cut scores change from the first year to the last. Table 2 shows that state test scores 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 one out of three state tests is unchanged through a matched NAEP era.

In some cases, in spite of comparable scales, EDFacts 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) or because participation rates in state test scores are below 95%. These estimable state trends and their NAEP counterparts are the data to which we fit our modeling framework in this next section.

State and NAEP Trend Modeling Framework

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

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

$$\boldsymbol{\delta}_s = \begin{bmatrix} \delta_{Ns} \\ \delta_{Ts} \end{bmatrix} \sim \left[\begin{pmatrix} \gamma_N \\ \gamma_T \end{pmatrix}, \begin{pmatrix} \tau_0 & \tau_{01} \\ \tau_{01} & \tau_1 \end{pmatrix} \right] = [\boldsymbol{\Gamma}, \boldsymbol{\tau}] \tag{1}$$

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:

$$\boldsymbol{e}_s \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{Ns} & 0 \\ 0 & \sigma_{Ts} \end{pmatrix} \right] = (\mathbf{0}, \boldsymbol{\sigma}_s). \tag{2}$$

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

changes is:

$$\begin{bmatrix}\boldsymbol{\delta}_s\\ \widehat{\boldsymbol{\delta}}_s\end{bmatrix}=\begin{bmatrix}\delta_{Ns}\\ \delta_{Ts}\\ \hat{\delta}_{Ns}\\ \hat{\delta}_{Ts}\end{bmatrix}\sim\left[\begin{pmatrix}\gamma_N\\ \gamma_T\\ \gamma_N\\ \gamma_T\end{pmatrix},\begin{pmatrix}\tau_0 & \tau_{01} & \tau_0 & \tau_{01}\\ \tau_{01} & \tau_1 & \tau_{01} & \tau_1\\ \tau_0 & \tau_{01} & \tau_0+\sigma_{Ns} & \tau_{01}\\ \tau_{01} & \tau_1 & \tau_{01} & \tau_1+\sigma_{Ts}\end{pmatrix}\right]=\left[\begin{pmatrix}\boldsymbol{\Gamma}\\ \boldsymbol{\Gamma}\end{pmatrix},\begin{pmatrix}\boldsymbol{\tau} & \boldsymbol{\tau}\\ \boldsymbol{\tau} & \boldsymbol{\tau}+\boldsymbol{\sigma}_s\end{pmatrix}\right] \tag{3}$$

This joint distribution implies the following data generating model:

$$\begin{aligned}&\hat{\delta}_{ts}=\delta_{Ns}(N)+\delta_{Ts}(T)+e_{ts}\\ &\delta_{Ns}=\gamma_N+u_{Ns}\\ &\delta_{Ts}=\gamma_T+u_{Ts}\\ &e_{ts}\sim(0,\sigma_{ts});[u_{Ns},u_{Ts}]'=\mathbf{u}_s\sim(\mathbf{0},\boldsymbol{\tau})\end{aligned} \tag{4a}$$

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,

$$\begin{aligned}&\widehat{\boldsymbol{\delta}}_s=\boldsymbol{\Gamma}+\mathbf{u}_s+\mathbf{e}_s;\\ &\mathbf{e}_s\sim(\mathbf{0},\boldsymbol{\sigma}_s),\mathbf{u}_s\sim(\mathbf{0},\boldsymbol{\tau})\end{aligned} \tag{4b}$$

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 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. 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 subject-grade-year changes and variances/covariances.

Fitting model (4) requires that we know the 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 error variance due to linking error or discretization error in state tests. For NAEP, linking 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, both linking and discretization error 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 these sources of error are negligible and assume the HETOP standard errors capture all sources of error variance. This approach will almost certainly 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 sensitivity analyses, we add additional error variance to the state test score estimates to identify a plausible range of estimated

correlations. For NAEP, linking error is assumed present but negligible relative to sampling error given the robust matrix sampling design employed (Mislevy et al., 1992); we therefore treat the NAEP-reported standard errors as capturing all error variance components.

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, $\hat{\delta}_{Ns}$ 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}:

$$
\begin{aligned}
bias(\hat{\delta}_{Ns}) &= E[\hat{\delta}_{Ns} - \delta_{Ns}] = 0; \\
MSE(\hat{\delta}_{Ns}) &= E\left[(\hat{\delta}_{Ns} - \delta_{Ns})^2\right] = \sigma_{Ns} \\
bias(\hat{\delta}_{Ts}) &= E[\hat{\delta}_{Ts} - \delta_{Ts}] = 0; \\
MSE(\hat{\delta}_{Ts}) &= E\left[(\hat{\delta}_{Ts} - \delta_{Ts})^2\right] = \sigma_{Ts}
\end{aligned}
\tag{5}
$$

Second, each of the observed test score changes—$\hat{\delta}_{Ts}$ and $\hat{\delta}_{Ns}$—provide 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:

$$
\begin{aligned}
&bias\left(\hat{\delta}_{Ts}\right) = E\left[\hat{\delta}_{Ts} - \delta_{Ns}\right] = \gamma_T - \gamma_N = b;\\
&MSE\left(\hat{\delta}_{Ts}\right) = E\left[\left(\hat{\delta}_{Ts} - \delta_{Ns}\right)^2\right] = b^2 + \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.\\
&bias\left(\hat{\delta}_{Ns}\right) = E\left[\hat{\delta}_{Ns} - \delta_{Ts}\right] = \gamma_N - \gamma_T;\\
&MSE\left(\hat{\delta}_{Ns}\right) = E\left[\left(\hat{\delta}_{Ns} - \delta_{Ts}\right)^2\right] = b^2 + \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ns}
\end{aligned}
\tag{6}
$$

Third, if the bias of the estimators above are 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, with

$$
\begin{aligned}
&bias\left(\hat{\delta}'_{Ts}\right) = E\left[\hat{\delta}_{Ts} - b - \delta_{Ns}\right] = 0;\\
&MSE\left(\hat{\delta}'_{Ts}\right) = E\left[\left(\hat{\delta}_{Ts} - b - \delta_{Ns}\right)^2\right] = \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ts}.\\
&bias\left(\hat{\delta}'_{Ns}\right) = E\left[\hat{\delta}_{Ns} + b - \delta_{Ts}\right] = 0;\\
&MSE\left(\hat{\delta}'_{Ns}\right) = E\left[\left(\hat{\delta}_{Ns} + b - \delta_{Ts}\right)^2\right] = \tau_0 + \tau_1 - 2\tau_{01} + \sigma_{Ns}.
\end{aligned}
\tag{7}
$$

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

$$\delta^*_{Ts} = E\left[\delta_{Ts}\middle|\hat{\delta}_{Ns}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ns}\right] = \gamma_T + \frac{\tau_{01}}{\tau_0 + \sigma_{Ns}}\left(\hat{\delta}_{Ns} - \gamma_N\right)$$

$$v^*_{Ts} = Var\left[\delta_{Ts}\middle|\hat{\delta}_{Ns}, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \sigma_{Ns}\right] = \tau_1(1 - \lambda_{Ns}r^2) \tag{9}$$

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

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

The MSE of these estimators is given by their posterior variances, v^*_{Ns} and v^*_{Ts}, respectively.

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

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

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

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

$$v_{1s}^* = Var\left(\delta_{Ts}\middle|\widehat{\boldsymbol{\delta}}_s, \boldsymbol{\Gamma}, \boldsymbol{\tau}, \boldsymbol{\sigma}_s\right) = \tau_1(1 - \lambda_{Ns}r^2)\left(\frac{1 - \lambda_{Ts}}{1 - \lambda_{Ns}\lambda_{Ts}r^2}\right)$$

(12)

We can compare the MSE of the estimators:

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

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

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

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

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

(13)

It is straightforward to show that

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

and

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

(14)

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 $\boldsymbol{\Gamma}$; $\hat{\delta}^*_{Ns}$ relies on $\hat{\delta}_{Ts}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\tau}$, and σ_{Ts}; $\hat{\delta}^*_{0s}$ relies on $\widehat{\boldsymbol{\delta}}_s$, $\boldsymbol{\Gamma}$, $\boldsymbol{\tau}$,

and $\boldsymbol{\sigma}_s$. In each case, using additional prior information sharpens our estimate of δ_{Ns}.[5]

The inequalities in (14) 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 $\hat{\delta}_{Ns}$. Once we know $\boldsymbol{\Gamma}$ 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 paired NAEP and state-test 2-year changes in math and reading for grades 4 and 8. We assume no linking and discretization error in our initial specification. We compare two analytic samples, one consisting of all year-pairs with valid state trends 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. Table 4

[5] The middle inequality is proven here:

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

The equality holds *iff*

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

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}. Table 4 reports RMSE, the square root of MSE in the prior section.

We briefly discuss here eight substantively important patterns in Table 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 could 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). It may also reflect linking or discretization error variance that the model does not account for.

Second, the estimated true correlation between NAEP and state-test changes is moderate and far from unity. In the full specification, the correlation is 0.47, and in recent years, the correlation is 0.61 (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.

Third, NAEP changes are markedly less reliable than state-test changes. In pooled specifications, NAEP reliabilities are 0.49 overall and 0.38 in recent years, and in specific year-pairs they range from 0.17 in 2022-2024 to 0.72 in 2019-2022. In contrast, 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.

Fourth, state assessments tend to report more positive changes than NAEP. In the pooled specification, the average state-NAEP discrepancy ("bias") is 0.045 SD units overall and 0.049 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

Table 4. Model parameter estimates for state and NAEP trends, 2009-2024, pooled by year models, assuming no linking or discretization 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)	0.00101	0.00127	0.00106	0.00223	0.00093	0.00112	0.00139	0.00329	0.00046
Var(State change)	0.00746	0.00784	0.01042	0.01077	0.00962	0.00518	0.00955	0.01099	0.00842
Cov(NAEP, State)	0.00168	0.00148	0.00150	0.00171	0.00062	0.00086	0.00278	0.00406	0.00137
True Correlation	0.61	0.47	0.45	0.35	0.21	0.36	0.76	0.68	0.70
Reliabilities									
NAEP	0.38	0.49	0.38	0.60	0.32	0.38	0.46	0.62	0.16
State	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
Average Change									
NAEP	-0.008	0.014	0.038	0.045	-0.001	-0.007	-0.018	-0.126	-0.002
State	0.041	0.060	0.094	0.043	0.084	0.033	0.038	-0.154	0.050
Bias									
NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State	0.049	0.045	0.056	-0.001	0.085	0.040	0.055	-0.028	0.052
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
EB (using state)	0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000
SD(State*)	0.020	0.019	0.015	0.016	0.007	0.012	0.029	0.041	0.015
Correlation with True NAEP Change									
NAEP	0.57	0.63	0.60	0.77	0.56	0.60	0.67	0.78	0.39
State	0.61	0.47	0.45	0.35	0.21	0.35	0.76	0.67	0.69
State - Bias	0.61	0.47	0.45	0.35	0.21	0.35	0.76	0.67	0.69
EB (using state)	0.61	0.47	0.45	0.35	0.21	0.35	0.76	0.67	0.69
RMSE									
NAEP	0.046	0.044	0.044	0.040	0.045	0.044	0.041	0.046	0.050
State	0.087	0.091	0.108	0.099	0.129	0.079	0.092	0.074	0.095
State - Bias	0.072	0.079	0.093	0.099	0.097	0.069	0.074	0.079	0.079
EB (using state)	0.025	0.032	0.029	0.044	0.030	0.031	0.024	0.042	0.015
EB (using NAEP & state)	0.022	0.026	0.024	0.030	0.025	0.026	0.021	0.031	0.015
N obs	654	1,256	266	284	52	210	186	244	258
Npairs	327	628	133	142	26	105	93	122	129
Pairs Included									
2009-11		x	x						
2011-13		x		x					
2013-15		x			x				
2015-17	x	x				x			
2017-19	x	x					x		
2019-22								x	
2022-24	x	x							x
Math	x	x	x	x	x	x	x	x	x
Reading	x	x	x	x	x	x	x	x	x
Grade 4	x	x	x	x	x	x	x	x	x
Grade 8	x	x	x	x	x	x	x	x	x
All Obs in Era	x	x	x	x	x	x	x	x	x
Controls Included									
Subject-Grade-Year FEs	x	x							
Subject-Grade FEs			x	x	x	x	x	x	x
Year FEs									

specific causal mechanisms.

Fifth, the relationships among bias, correlation, and precision clarify why state trends can be simultaneously "biased" yet also informative about true NAEP change. Despite the positive mean discrepancy, the observed state-test changes remain correlated with true NAEP changes because (i) the true correlation ρ_{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 4 quantify the consequences for estimation of δ_{Ns}. Following Equation 12, 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 minimal 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 $\rho_{NT} < 1$, 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 5 shows pooled results separately by subject and grade, for 2009-2024 and 2015-2024, both excluding the 2019-2022 pandemic, 3-year 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

Table 5. Model parameter estimates with pooled models by subject and grade, assuming no linking or discretization 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)	0.00166	0.00119	0.00085	0.00038	0.00194	0.00091	0.00157	0.00079
Var(State change)	0.00573	0.00953	0.00563	0.00851	0.00640	0.00982	0.00596	0.00870
Cov(NAEP, State)	0.00221	0.00223	0.00138	0.00075	0.00204	0.00110	0.00187	0.00098
True Correlation	0.72	0.66	0.63	0.41	0.58	0.37	0.61	0.38
Reliabilities								
NAEP	0.47	0.41	0.36	0.21	0.56	0.41	0.57	0.40
State	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.99
Average Change								
NAEP	0.027	-0.024	-0.004	-0.030	0.034	0.007	0.010	0.007
State	0.065	0.037	0.040	0.023	0.075	0.053	0.054	0.056
Bias								
NAEP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
State	0.038	0.061	0.044	0.053	0.040	0.046	0.044	0.050
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
EB (using state)	0.000	0.000	0.000	0.000	-0.003	-0.001	0.000	0.000
SD(State*)	0.031	0.023	0.018	0.008	0.029	0.012	0.026	0.013
Correlation with True NAEP Change								
NAEP	0.64	0.59	0.56	0.41	0.69	0.56	0.69	0.55
State	0.71	0.66	0.62	0.41	0.58	0.36	0.60	0.37
State - Bias	0.71	0.66	0.62	0.41	0.58	0.36	0.60	0.37
EB (using state)	0.71	0.66	0.62	0.41	0.58	0.36	0.60	0.37
RMSE								
NAEP	0.048	0.047	0.043	0.044	0.046	0.045	0.041	0.043
State	0.067	0.100	0.076	0.101	0.077	0.104	0.076	0.100
State - Bias	0.055	0.080	0.062	0.087	0.066	0.093	0.063	0.087
EB (using state)	0.029	0.026	0.023	0.018	0.036	0.028	0.032	0.026
EB (using NAEP & state)	0.025	0.023	0.020	0.016	0.028	0.024	0.025	0.022
N obs	184	182	126	162	338	340	254	324
Npairs	92	91	63	81	169	170	127	162
Pairs Included								
2009-11	x	x	x	x	x	x	x	x
2011-13	x	x	x	x	x	x	x	x
2013-15	x	x	x	x	x	x	x	x
2015-17	x	x	x	x	x	x	x	x
2017-19	x	x	x	x	x	x	x	x
2019-22								
2022-24	x	x	x	x	x	x	x	x
Math	x		x		x		x	
Reading		x		x		x		x
Grade 4	x	x			x	x		
Grade 8			x	x			x	x
All Obs in Era	x	x	x	x	x	x	x	x
Controls Included								
Subject-Grade-Year FEs								
Subject-Grade FEs								
Year FEs	x	x	x	x	x	x	x	x

between NAEP Reading and state tests of English Language Arts, which sometimes incorporate information about progress in other domains like Writing, might cause disagreements in trends.

Eighth and finally, the expanded 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. This is modest in magnitude because state-test reliability is already near one, but it is nontrivial and directionally consistent with the theory that conditioning on both signals cannot increase MSE. This estimator is particularly relevant for "smoothing" historical NAEP trends in windows where both NAEP and state assessments are available. In contrast, the state-only shrinkage estimator is most relevant for "predicting" NAEP-scale changes in windows where NAEP is unavailable but state testing continues within a stable regime.

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, once we account for the variance components and reliabilities implied by Model 1, 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.

Secondary Analyses

We test theoretical and practical extensions of our modeling framework with two analyses in this section:

(1) a sensitivity analysis adding putative linking/discretization error variance to state-test estimates, and (2) an out-of-sample prediction exercise. For the first analysis, as described above, our primary specification treats state-test measurement-error variances as known from HETOP estimation. This primarily captures uncertainty from student sampling and explains the finding that state test trend reliability is estimated near unity. We test the sensitivity of our results to two additional sources of error likely to be present in HETOP-estimated test score trends: linking error and discretization error.

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 do not find these estimates in technical reports for state testing programs. To evaluate the robustness of our results to underestimation of state trend measurement error, we use estimates from Michaelides and Haertel (2014). They use a bootstrap approach to estimate linking error variance on the order of 0.002 in standardized units per linking operation for a state test with 44 common items embedded across 8 matrix-sampled forms. Because a 2-year change involves two such links, and assuming linking errors across adjacent years are approximately uncorrelated, a plausible approximation of linking error variance for a test like this is 0.002–0.004 per 2-year change.

Table 6 presents a sensitivity analysis in which we add fixed amounts of error variance (0.00025–0.004) to the state test change measurement-error variances before fitting the model. The results show that state reliability falls substantially as assumed linking error increases, from 0.99 in the base case to 0.48 under the most extreme assumption. The estimated true correlation rises correspondingly (from 0.61 to 0.87), reflecting disattenuation. Critically, however, estimated bias, RMSE comparisons across estimators, and shrinkage estimates themselves are unchanged across all specifications. Every estimator of the true NAEP change and its MSE depends on τ_1 and σ_T^2 only through their sum, the total observed variance of state test score trends. Adding linking/discretization error variance λ 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.

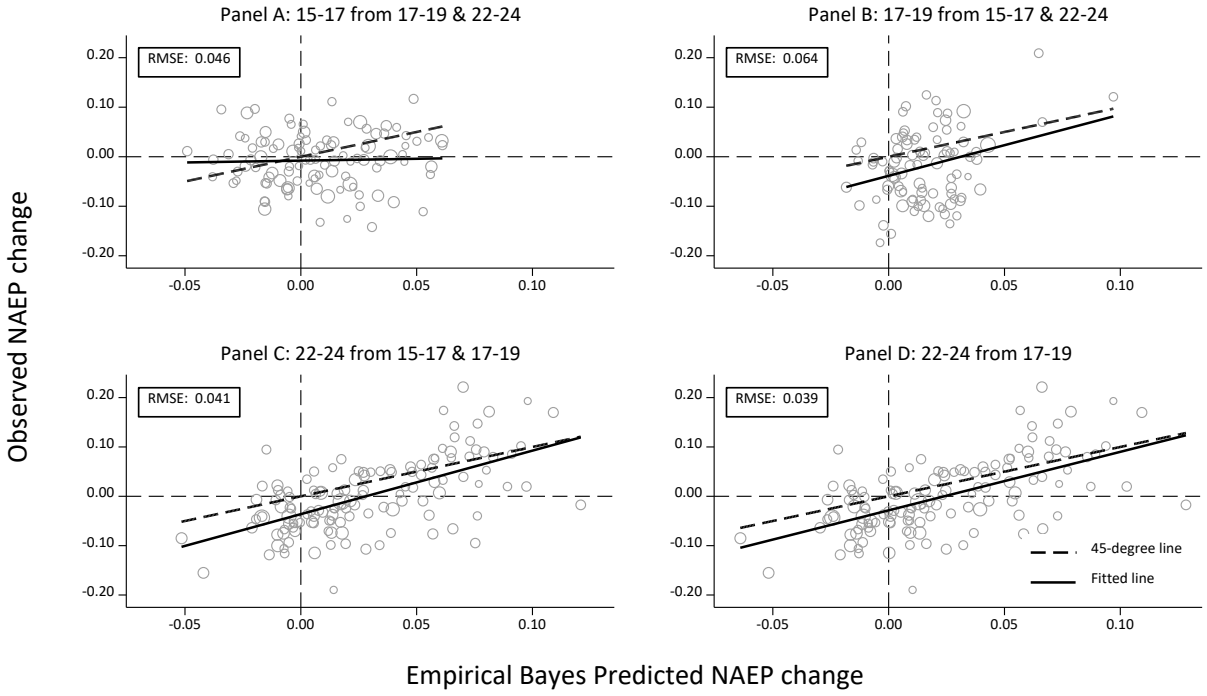
Table 6. Sensitivity of model estimates to underestimated state test score trend imprecision.

	Original	1	2	3	4	5
Added Variance	0	0.00025	0.0005	0.001	0.002	0.004
Tau Matrix						
Var(NAEP change)	0.00101	0.00101	0.00100	0.00100	0.00100	0.00100
Var(State change)	0.00746	0.00736	0.00714	0.00667	0.00568	0.00369
Cov(NAEP, State)	0.00168	0.00169	0.00169	0.00168	0.00168	0.00168
True Correlation	0.61	0.62	0.63	0.65	0.70	0.87
Reliabilities						
NAEP	0.38	0.38	0.38	0.38	0.38	0.38
State	0.99	0.96	0.92	0.86	0.73	0.48
Average Change						
NAEP	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008
State	0.041	0.040	0.040	0.040	0.040	0.040
Bias						
NAEP	0.000	0.000	0.000	0.000	0.000	0.000
State	0.049	0.048	0.048	0.048	0.048	0.048
State - Bias	0.000	0.000	0.000	0.000	0.000	0.000
EB (using state)	0.000	0.000	0.000	0.000	0.000	0.000
SD(State*)	0.020	0.020	0.020	0.020	0.020	0.020
Correlation with True NAEP Change						
NAEP	0.57	0.57	0.57	0.57	0.57	0.57
State	0.61	0.61	0.60	0.60	0.60	0.60
State - Bias	0.61	0.61	0.60	0.60	0.60	0.60
EB (using state)	0.61	0.61	0.60	0.60	0.60	0.60
RMSE						
NAEP	0.046	0.046	0.046	0.046	0.046	0.046
State	0.087	0.087	0.088	0.088	0.088	0.088
State - Bias	0.072	0.073	0.073	0.074	0.074	0.074
EB (using state)	0.025	0.025	0.025	0.025	0.025	0.025
EB (using NAEP & state)	0.022	0.022	0.022	0.022	0.022	0.022
N obs	654	654	654	654	654	654
Npairs	327	327	327	327	327	327

We also evaluate how well the state-only shrinkage estimator predicts NAEP trends out of sample, in year-pairs whose data were not used to fit the model. This out-of-sample evaluation simulates the situation we face in predicting 2024-25 NAEP trends from 2024-25 state test data in the absence of a 2025-NAEP administration. 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 0.041 (Panel C, trained on 2015-17 and 2017-19) and 0.039 (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, consistent with the lower true NAEP-state trend correlation estimated in that earlier period (Table 4). Predicting 2015-17 NAEP changes from later years is hardest (Panel A, RMSE = 0.046), consistent with state trends being less informative about NAEP trends in that period. These RMSEs compare favorably to two-year NAEP-only RMSEs from Table 4, suggesting that state-only shrinkage estimators can match or exceed the precision of NAEP alone in held-out years. 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.

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



Discussion and Conclusion

This paper addresses two research questions that inform educational progress monitoring in the United States. First, we estimate the relationship between state test score trends and NAEP trends from 2009-2024. We find that NAEP trends are only moderately reliable. The two trends are correlated roughly 0.5-0.6 at minimum. Incorporating linking and/or discretization error in state test trends would increase this estimate. We encourage further research and documentation of the degree of state linking error for trend estimation to better inform estimates of state test score trend reliability and true correlations between state test score trend and NAEP trends. State tests are more positive, on average, than NAEP trends by approximately 0.04-0.05 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). Second, we show that this model

implies properties of five different estimates of true NAEP trends. State trends are poor and biased estimates of NAEP trends, but Empirical Bayes shrinkage estimators that exploit the state test's high precision and moderate alignment with NAEP can deliver lower mean squared error than NAEP trend estimates alone. We use these estimates to show that state test scores can support predictions of NAEP trends in the absence of NAEP data when comparable state test score data exist.

Methodologically, we treat discrepancies between the two testing systems not as a problem to resolve but as structured information to synthesize. By decomposing disagreement into (i) measurement error, (ii) systematic mean bias, and (iii) variance in true trend discrepancies, our framework provides a family of estimators whose MSE can be compared analytically from a single fitted model. This decomposition clarifies why neither test alone is optimal. NAEP provides an unbiased but imprecise signal. State tests provide a possibly precise but biased signal. The optimal estimator, when both signals are available, is the two-signal Empirical Bayes posterior mean, which strictly dominates any single-signal estimator in MSE. We show how this estimate performs out-of-sample to show that the modeled RMSE advantage can translate to real predictive improvement. Future extensions can model the causes and timing of state trend breaks, allow bias and alignment to vary across states and eras, and incorporate additional tests to further improve mapping of the patchwork landscape of educational progress data 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.

Braun, H., Zhang, J., & Vezzu, S. (2010). An investigation of bias in reports of the National Assessment of Educational Progress. *Educational Evaluation and Policy Analysis, 32*, 24–43. https://doi.org/10.3102/0162373709351137

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: 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., & 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

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.

McLaughlin, D. (2005). *Properties of NAEP full population estimates* (Unpublished report). American Institutes for Research. Retrieved from http://www.schooldata.org/Portals/0/uploads/reports/NSA_T1.5_FPE_Report_090205.pdf

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.

National Institute of Statistical Sciences. (2009). *NISS/NESSI task force on full population estimates for NAEP* (Technical Report #172). Retrieved from http://www.niss.org/sites/default/files/technical_reports/tr172.pdf

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