Validation Methods for Aggregate-Level Test Scale Linking: A Rejoinder

Validation methods for aggregate-level test scale linking: A rejoinder

Andrew D. Ho, *Harvard Graduate School of Education*
Sean F. Reardon, *Stanford University*
Demetra Kalogrides, *Stanford University*

In Reardon, Kalogrides, and Ho (2021), we developed precision-adjusted random effects models to estimate aggregate-level linking error, for populations and subpopulations, for averages and progress over time. We are grateful to past editor Dan McCaffrey for selecting our paper as the focal article for a set of commentaries from our colleagues, Daniel Bolt, Mark Davison, Alina von Davier, Tim Moses, and Neil Dorans. These commentaries reinforce important cautions and identify promising directions for future research. In this rejoinder, we clarify aspects of our originally proposed method. 1) Validation methods provide evidence of benefits and risks that different experts may weigh differently for different purposes. 2) Our proposed method differs from "standard mapping" procedures using the National Assessment of Educational Progress not only by using a linear (vs. equipercentile) link but also by targeting direct validity evidence about counterfactual aggregate scores. 3) Multilevel approaches that assume common score scales across states are indeed a promising next step for validation, and we hope that states enable researchers to use more of their common-core-era consortium test data for this purpose. Finally, we apply our linking method to an extended panel of data from 2009 to 2017 to show that linking recovery has remained stable.

Validation methods for aggregate-level test scale linking: A rejoinder

In Reardon, Kalogrides, and Ho (2021), we developed precision-adjusted random effects models to estimate aggregate-level linking error, for populations and subpopulations, for averages and progress over time. We are grateful to past editor Dan McCaffrey for selecting our paper as the focal article for a set of commentaries from our colleagues, Daniel Bolt, Mark Davison, Alina von Davier, Tim Moses, and Neil Dorans. These commentaries reinforce important cautions and identify promising directions for future research. In this rejoinder, we clarify aspects of our originally proposed method. 1) Validation methods provide evidence of benefits and risks that different experts may weigh differently for different purposes. 2) Our proposed method differs from "standard mapping" procedures using the National Assessment of Educational Progress not only by using a linear (vs. equipercentile) link but also by targeting direct validity evidence about counterfactual aggregate scores. 3) Multilevel approaches that assume common score scales across states are indeed a promising next step for validation, and we hope that states enable researchers to use more of their common-core-era consortium test data for this purpose. Finally, we apply our linking method to an extended panel of data from 2009 to 2017 to show that linking recovery has remained stable.

The federal ED*Facts* database holds approximately 500 million coarsened, aggregated, student test scores from U.S. public school students in grades 3-8. Tests differ across states and change over time, so these data sit in silos, incomparable across states and years. The burden of a coherent story about national educational progress falls to the National Assessment of Educational Progress (NAEP), where reported results are limited to states and large urban districts. Researchers interested in comparing educational progress and growth at the district level enter a metaphorical Tower of Babel. The contribution of our original paper (Reardon et al., 2021) was not to recognize that NAEP could serve as a

Rosetta Stone (others had already recognized that possibility[1]), but to recognize that direct validity evidence for the accuracy of the "translations" of linked scores can sometimes hide in plain sight. In this case, the targets for recovery are NAEP results for urban districts. We show how precision-adjusted random effects models can summarize the recovery of these parameters. The degree of recovery should guide the use of linked scores.

In their four responses, our colleagues generally agree that this kind of validity evidence is useful. Where we disagree, it is less with the method we propose than the conclusions we draw from the application of our methods to our illustrative case study. In our original paper, we concluded that, "the linked estimates are accurate enough to be used to investigate broad patterns in the relationships between average test performance and local community or schooling conditions, both within and between states" (p. TBD). Our colleagues are more cautious, using descriptors like "preliminary support" (Bolt, this issue, p. TBD); "substantial, but incomplete, support" (Davison, this issue, p. TBD); "the effects of this bias need to be further investigated" (von Davier, this issue, p. TBD); and "we urge cautious use of the results" (Moses & Dorans, this issue, p. TBD).

We respect their cautions and believe that we understand their perspective. In the first part of our rejoinder, we explain why think we largely agree with our colleagues about the evidence and what the evidence means. We then explain why we nonetheless stand behind our conclusion, because we believe our colleagues are underestimating the benefits and overestimating the potential for downstream harm. In the second and third parts of our rejoinders, we discuss two technical points that our colleagues

[1] The appealing possibility that NAEP could "translate" state and local tests onto a common scale was an explicit motivation behind the development of state-level NAEP, as is clear from the Alexander-James Report that launched NAEP's modern era, "we recommend that the national assessment devise a linkage system relating local and state testing and assessment programs to the national assessment" (Alexander & James, 1987, p. 13). Thissen (2007) also provides a history of aggregate linking efforts.

raised in common across most of their responses, about the NAEP-state standard mapping procedure and multilevel linking approaches, respectively. In the fourth part, we address a few specific residual points in each of the responses. In the fifth and final part of our rejoinder, we revisit our linking validation study with the benefit of two additional waves of NAEP data, in 2015 and 2017. The results show that the linking performance is stable.

The appropriate use of extant test score data for aggregate-level research

Given the orientation of this journal toward methods, our original paper focused more on our approach and less on guidance for the appropriate use of linked scores from our specific case study. We nonetheless stated our intended and unintended uses explicitly: "enabling large-scale research about educational achievement" (p. TBD) and, "we do not attempt to estimate student-level scores, and we do not intend the results to be used for high-stakes accountability" (p. TBD). These statements are important. Random errors that are reduced at the aggregate level would be substantial for individual students.[2] And placing high stakes on linked test scores from tests that measure different content could degrade future linked results.

We are glad to see our colleagues echoing these cautions. However, we find the balance of some of their remarks to overemphasize the potential harm of linked scores without acknowledging actual benefits. For example, von Davier (this issue) states, "the situations in which the use of this approximate

[2] Figure 3 from our original paper shows the relationship between the reliability of linked means on the NAEP scale in terms of linking error and the standard error of district average scores on state tests (which is a function of district size). Assuming linking error is the same as that for TUDAs, we show that reliabilities are above .70 for 91% of districts, almost all but the smallest school districts. This illustrates how aggregation improves the precision of comparisons. This is consistent with standard examples from measurement that show that correlations are stronger, and residuals are smaller for aggregates than they are for individual students.

linking is not damaging to the districts may be limited" (p. TBD). Davison (this issue) remarks, "once the linking function is established, the barn door is wide open" (p. TBD). While we can imagine damaging uses, it is a "slippery slope" argument to discourage reported scores that have actual beneficial uses based only on possible misuse and harm. We should be specific. Possible misuse is not misuse.

Of course, our colleagues have decades of experience seeing folks slide down slippery slopes. In their article defining "Intuitive Test Theory," Braun and Mislevy (2005) describe the relevant, intuitive, and generally false assumption: "any two tests that measure the same thing can be made interchangeable, with a little 'equating magic'" (p. 493). Why does this intuition generally fail? Tests that purport to measure the same thing rarely do. And competing tests typically serve different purposes, leading to degraded linking functions over time. Haertel and Ho (2016) describe linkages of dissimilar tests as a "construct shift," where an interpretation or use of linked scores differs from those of originally reported scores and relies on evidence that the original test developers never imagined nor sought to collect.

The antidotes to false intuition and unsupported shifts are clear statements of intended uses and evidence supporting these uses. Rather than rely on "equating magic" or fail to imagine "construct shifts," we believe our article presents a useful method for estimating bias and error in linked scores. And we believe our case study makes our intended uses and interpretations of linked scores clear, as well as our evaluation of the evidence for these intended uses.

What are the benefits? We believe that aggregate test score data from state testing programs are a public good. They should be part of a research infrastructure. The 500 million test scores in the aforementioned ED*Facts* database represent billions of dollars of taxpayer investment and weeks worth

of time each year spent by every public school student, teacher, and administrator in this country. Rather than assume this investment should remain locked in silos by state and year, our method evaluates whether comparisons across states and years are defensible in aggregate, to enable descriptive and causal research about educational opportunity. There is certainly possible harm of a user misinterpreting any given district-to-district comparison despite our cautions and caveats. This possible harm must be balanced against the benefit of research that leverages linked data to discover new descriptive and causal patterns that can improve our understanding and implementation of education.

The relationship between our approach and NAEP-state standard mapping

All four responses contrasted our approach with the NAEP-state standard mapping published regularly by the National Center of Education Statistics (e.g., Bandeira de Mello, Rahman, Fox, & Ji, 2019) and estimated using methods from McLaughlin and Bandeira de Mello (2006) and Braun and Qian (2007). Davison (this issue) observes correctly that our approach is linear whereas theirs is equipercentile. Moses and Dorans (this issue) argue that critiques of the NAEP-state standard mapping (e.g., Ho & Haertel, 2007; Koretz, 2007) should also apply to our approach.

Our approaches differ in method, quality, and purpose. Our approach evaluates whether linked NAEP urban district average scores recover actual NAEP urban district average scores. In this way, we evaluate the recovery of a subset of target parameters directly. Our bias and RMSE statistics are direct measures of our (in)ability to answer the essential counterfactual question, "what would districts score had they taken NAEP?" In contrast, the NAEP-state standard mapping attempts to recover a state-level performance standard on an alternative scale. The counterfactual question is, "where would states have set a cut score if NAEP were their state test?" As Ho and Haertel (2007) observe in their critique, the

Braun and Qian (2007) approach sets this counterfactual standard tautologically, regardless of the relationship between NAEP and state test scores. It is not testable directly. Instead, the mapping procedure uses correlations of school-level proficiency percentages to support the concordance of state and NAEP results. High correlations support the validity of the mapped standard. However, such evidence is indirect, whereas ours is direct.

The equipercentile linking approach underlying the NAEP-state standard mapping is particularly appropriate for its intended purpose of recovering a cut score that sets a given percentage of students as "proficient." We could also estimate and apply an equipercentile linking function as Davison (this issue) suggests. This would entail 1) establishing a mixture distribution of estimated district distributions in each state, 2) estimating an equipercentile linking function between this mixture distribution and NAEP plausible values in each state, 3) transforming state test scores using this function, and 4) re-estimating means and standard deviations of transformed state test scores for each district. This requires evidence or accurate assumptions about percentiles of state test score distributions.

We use data from ED*Facts* where states report frequencies in "coarsened" proficiency categories. Our past research has shown that we can recover means and standard deviations from such data (Reardon, Shear, Castellano, & Ho, 2017). However, we have less confidence in the recovery of percentiles, particularly at extremes of state test score distributions. Absent such evidence, we opt for a linear linking function. However, we agree with Davison (this issue) that we can use uncoarsened state test score distributions to compare linear and equipercentile linking functions. Overestimation of urban district average scores could theoretically be ameliorated by a nonlinear linking function that emphasizes lower ranges of state test score scales compared to NAEP.

The promise of multilevel models for comparing state linking functions

Bolt (this issue), Davison (this issue), and Moses and Dorans (this issue) all raise promising possibilities for evaluating differential linking relationships across states. When states do not share test score scales, we can estimate relationships among scores within each state and compare these relationships across states. When states do share test score scales, our colleagues correctly note that differential prediction models can evaluate whether linkages are stable across states. These efforts enable us to distinguish within-state and between-state contributions to linking error, as Bolt (this issue) notes.

In our NAEP application, there are few districts and fewer states with multiple districts, thus sample sizes were too small to enable precise distinction between within-state and between-state linking relationships. However, extending our model and method to incorporate state membership is indeed a natural next step. If the NAEP program continues to expand the number of urban districts, this may prove to be fruitful in the future. We could also have fit such models to districts from our NWEA MAP example (e.g., Figure 2) given the large number of districts within many states. A future Data Use Agreement could allow us to investigate this. More states have come to share common assessments in recent years, although many purportedly identical assessments have imposed slight content differences to meet state alignment criteria. Nonetheless, analyses using these common assessments should be able to improve our understanding of sources of linking error within and between states.

Moses and Dorans (this issue) also present an analysis of SAT data. They use Evidence-based Reading and Writing (ERW) scores to estimate linking functions to enable the comparison of Math scores across states. Of course, SAT Math scores are already on a common scale, so, like us, Moses and Dorans can evaluate the recovery of district means directly. Because they have many districts within each state,

they can also evaluate linking error for each state. They use this analysis to make the point that seemingly high correlations do not preclude large deviations, particularly in some states.

Of course, we agree. This is why we reported both correlations and root mean square errors (RMSEs) in our Table 1. The results suggest that some number of districts may be misestimated slightly, and a smaller number may be misestimated substantially. Returning to our first point, this is why we do not recommend the use of our linking functions for individual scores or high-stakes accountability.

Responses to additional points from reviewers

Davison (this issue) describes a "regression to the mean" hypothesis, where lower scoring districts and groups appear to have more positive linking bias. We acknowledge the pattern, and we further believe it is important to distinguish between artifactual and substantive explanations for this pattern. An example of a substantive explanation could be where students in urban districts are also those who focus more on state (vs. NAEP) tested content than other districts. Note that this cannot apply to all low-scoring students or low-scoring districts. Because the linear linking enforces identical variance across tests, it is impossible to observe uniform regression to the mean across all districts. An example of artifactual regression to the mean could be where state test scores have less measurement error than their reliability statistics suggest or, more realistically, if NAEP corrections for measurement error are not sufficient. This could systematically shrink district averages and lead to conditional bias resembling "regression to the mean." Substantive explanations help us to understand the magnitude of linking errors but do not necessarily help us to reduce them. Artifactual explanations may indicate that we have misestimated the magnitude of the linking errors, in this case, by overestimating them.

Bolt (this issue) recommends analyses that help us to understand sources of bias, by predicting residuals using various district features. We agree that this would be an interesting effort even if, as we note above, it may not help us improve the linking itself. A predictive linking, where we adjust linked scores based on demographic variables to reduce linking error, could be undesirable for its asymmetry as well as its confounding of inferences about achievement and demographic characteristics.

von Davier comments that our paper, "seems to suggest differences across tests can be minimized despite differences in content, constructs, and response processes" (p. TBD). We agree that these could be consequential differences, as we noted on page TBD. Our method enables an empirical test of the impact of these consequences. Moses and Dorans describe our approach as "indirect validation," but we would describe our results in Tables 1 and 3 as direct validation. Both von Davier (this issue) and Moses and Dorans (this issue) suggest we explore subgroup analyses. It was not clear to us whether they meant something besides the subgroup analyses we had presented at the bottom of Table 1 in our original paper.

Extending linking results through 5 waves of data

We present an updated version of Table 1 from our paper here. This table includes linking performance through 2 additional waves of data for a total of 5 waves from 2009 through 2017. The results are very similar to those in our original paper. The bias in estimated means was smaller in 2015 and 2017 than in prior years. Overall, across years, subjects, and subgroups, the RMSE was generally in the range 3.5 to 5.5 points (average 4.2). Biases was -.5 to 2.5 points (average 0.9) and correlations were 0.93-0.98 (average 0.97). Similar tables are available for pooled district scores relating to averages, year-to-year

progress, and grade-to-grade slopes. We will continue to monitor the recovery of NAEP TUDA means as additional waves of data become available.

Discussion

We are grateful to our colleagues for their thoughtful, critical, and constructive remarks on our original paper. In this rejoinder, we have clarified that our method evaluates the direct recovery of target parameters. This stands in contrast to NAEP-state standard mapping, which offers only indirect support for tautological mapped standards. If our direct validation approach showed perfect recovery, this would render possible reasons for linking errors moot for any population of districts to which these results are generalizable. Of course, the recovery is not perfect. This leaves the possibility of reasonable disagreement about whether linked aggregate scores are sufficiently accurate for particular purposes, weighing tradeoffs in benefits and harm. We appreciate the sound advice of our colleagues as they identify next steps for research that will solidify the evidence base and improve appropriate uses of linked aggregate scores.

Table 1: Recovery of NAEP TUDA means following state-level linkage of state test score distributions to the NAEP scale, Measurement Error Adjusted.

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2009	16	4.57	2.25	0.94
		2011	20	4.19	1.28	0.95
		2013	20	3.06	0.26	0.97
		2015	19	4.36	0.25	0.97
		2017	24	4.26	-0.37	0.96
	8	2009	16	3.81	1.22	0.92
		2011	20	2.70	0.55	0.96
		2013	20	3.85	1.66	0.92
		2015	19	2.52	-0.24	0.97
		2017	24	4.18	0.49	0.95
Math	4	2009	16	6.58	4.23	0.92
		2011	20	5.22	2.65	0.93
		2013	20	4.07	1.62	0.94
		2015	19	2.43	0.79	0.99
		2017	24	3.05	0.16	0.98
	8	2009	13	5.89	3.77	0.95
		2011	17	4.02	2.03	0.96
		2013	14	5.28	1.58	0.93
		2015	15	4.35	-3.08	0.97
		2017	24	4.75	-1.52	0.94
Average		2009	13-16	5.29	2.83	0.93
		2011	17-20	4.14	1.61	0.95
		2013	14-20	4.04	1.26	0.94
		2015	15-19	3.49	-0.43	0.98
		2017	24	4.11	-0.31	0.96
		Reading	16-24	3.82	0.67	0.97
		Math	13-24	4.60	1.07	0.96
		All	13-24	4.22	0.87	0.97
Subgroup Average		Male		4.71	1.16	0.97
		Female		4.37	1.36	0.97
		White		5.89	0.83	0.96
		Black		4.84	1.14	0.94
		Hispanic		5.08	2.11	0.96

Source: Authors calculations from ED*Facts* and NAEP TUDA Expanded Population Estimates data. Estimates are based on Equation 7 in Reardon et al. (this issue). Subgroup averages are computed from a model that pools across grades and years, like Equation 9 in Reardon et al. (this issue).

References

Alexander, L., & James, H. T. (1987). *The Nation's Report Card: Improving the Assessment of Student Achievement.* National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2018/02/Nations-Report-Card_Full.pdf

Bandeira de Mello, V., Rahman, T., Fox, M.A., and Ji, C.S. (2019). Mapping State Proficiency Standards onto the NAEP Scales: Results From the 2017 NAEP Reading and Mathematics Assessments (NCES 2019-040). U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2019040.pdf

Bolt, D. (2021). Commentary on Reardon, Kalogrides, and Ho. *Journal of Educational and Behavioral Statistics.*

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan,* 489-497.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, and P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 313-338).

Davison, M. (2021). Commentary on Reardon, Kalogrides, and Ho. *Journal of Educational and Behavioral Statistics.*

Haertel, E. H., & Ho, A. D. (2016). Fairness using derived scores. In N. Dorans and L. Cook (Eds.), *Fairness in educational assessment and measurement* (217-237). New York, NY: Routledge.

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

Moses, T., & Dorans, N. (2021). Aggregate-level test-scale linking: A new solution for an old problem? *Journal of Educational and Behavioral Statistics.*

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.*

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,* 3-45.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287-312). New York, NY: Springer.

von Davier, A. (2021). Commentary on Reardon, Kalogrides, and Ho. *Journal of Educational and Behavioral Statistics.*