

Apples to Apples? The Underlying Assumptions of State-NAEP Comparisons

Andrew Ho Edward Haertel
University of Iowa Stanford University

This policy brief is the second of a two-part series. The first, entitled, *(Over)-Interpreting Mappings of State Performance Standards onto the NAEP Scale*, anticipates an upcoming report from the National Center for Education Statistics (NCES). In this report, Professor Henry Braun of Boston College and Dr. Jiahe Qian of Educational Testing Service extend a method that locates state performance standards on the NAEP score scale. This mapping affords interpretations of state performance standards as “higher” or “lower” than other state or NAEP standards. Our first brief asserts that such interpretations depend logically on an argument for the equivalence of state tests and NAEP that is rarely advanced and probably not entirely valid. Reasonable equivalence between state tests and NAEP may or may not exist. However, we show that substantial differences between state tests and NAEP will render the mapping illogical and subject to drift over time.

This second brief makes three short points about continuing efforts to reference state tests to NAEP.

- 1) There is a large body of literature on linking state tests to NAEP; much of it is cautionary.**
- 2) Methods exist that can help to support or refute the validity of state mappings onto the NAEP scale.**
- 3) Perceptions of NAEP as an infallible “audit” test threaten current NAEP and state testing priorities.**

1) A Long History of State-NAEP Comparison

The concerns we raise in our first brief have been raised in similar forms time and time again. In late 1997, the National Research Council (NRC) formed a study committee to investigate using NAEP as a common scale for state tests. The idea arose as a response to President Clinton’s 1997 proposal for a Voluntary National Test (VNT) – could states maintain their own tests but report results on the NAEP score scale? The committee was asked “to determine if an equivalency scale can be developed that would allow test scores from commercially available standardized tests and state assessments to be compared with each other and the National Assessment of Educational Progress” (PL 105-78, SEC. 306 (a)).

As Linn (2005) describes in an overview of NAEP linking efforts, “the answer to the question presented to the committee was simply ‘no’” (p. 7). The resulting NRC report, entitled *Uncommon Measures*, includes a number of cautions that parallel ours from the first brief, for example, from the executive summary, “Links between most existing tests and NAEP... in terms of the NAEP achievement levels will be problematic. Unless the test to be linked to NAEP is

very similar to NAEP in content, format, and uses, the resulting linkage is likely to be unstable and potentially misleading” (Feuer, Holland, Green, Bertenthal & Hepmhill, 1999, p. 4).

The NRC report overviews a number of previous linking attempts. More recent literature has been reviewed by David Thissen’s chapter in an upcoming volume entitled *Linking and Aligning Scores and Scales* (Dorans, Pommerich, & Holland, in press). Braun and Qian also contribute a chapter to this volume and provide technical details about their method. Daniel Koretz’s commentary on the Braun and Qian method, also included in the volume, raises many of the same concerns we address in our first brief.

Achievement level comparisons between State and NAEP tests also predate NCLB. Mark Musick highlighted several State-State and State-NAEP discrepancies in 1996, leading Richard Hill (1997) to anticipate that increasing stakes on NAEP results may lead to distortions in NAEP results. The literature on achievement level reporting, meanwhile, stretches far back to the days of criterion-referenced testing and minimum competency. Gene Glass memorably described the process of setting performance standards as basing “a grand scheme on a fundamental, unsolved problem” (Glass, 1978, p. 237). While State-NAEP comparisons have become widespread under NCLB, concerns with their fundamental techniques and assumptions have been well detailed across various literatures for many years.

2) Methods for Evaluating the Alignment of State Tests and NAEP

Though it may sound like a platitude, one way to obtain an armchair estimate of the validity of a State-NAEP mapping is to look at State and NAEP items. Sample items from both State tests and from NAEP are available online, and in many cases the differences in format and approach are immediately apparent regardless of the psychometric training of the observer. These kinds of contrasts help to temper expectations about the similarity of results for professionals and laypersons alike.

Beyond this casual approach, there are numerous alignment theories and methodologies that advance understanding of relationships between content standards, tests, and teaching. A recent special issue of the journal *Applied Measurement in Education* overviews much of the literature on this topic (Herman & Webb, 2007). Articles by Norman Webb and Andrew Porter and his colleagues are particularly germane and describe well known alignment approaches that have been implemented in many states.

A great deal could be learned from an alignment study of state test items to NAEP frameworks using Webb’s approach, or state teaching practices to NAEP frameworks using Porter’s approach. Whether the results are quantitative or qualitative, it seems likely that alignment of state tests and state teaching practices to state content standards will be greater than their alignment to NAEP frameworks. As our companion brief explains, substantial differences in alignment render State-NAEP mappings illogical, like handicapping a golf game based on mini-golf performance.

The expected equivalence of State tests and NAEP is so pervasive that it is best characterized as being driven by an intuition. Braun and Mislevy (2005) have listed flawed but widespread

intuitions about testing in a *Phi Delta Kappan* article entitled, “Intuitive Test Theory.” The relevant intuitions underlying many State-NAEP comparisons are, “A test is a test is a test,” and, “A test measures what it says at the top of the page.” In other words, given one state and two “Reading” tests, why should results be different? The analyses described in the above references should help to dispel these intuitions.

3) Perceptions of NAEP as an infallible “audit” test are shortsighted.

When State-NAEP discrepancies arise, they are rarely treated with dispassion. Whether authors compare percents proficient, trends, or gaps, discrepancies are often framed as controversies where one result is valid and the other suspect. Some of these discrepancies are artifacts of the deeply flawed, percent proficient-based reporting scale (Holland, 2002; Ho and Haertel, 2005; Koretz and Hamilton, 2006), but discrepancies are manifest on better scales. A disappointing tendency has been to assume, in the absence of explicit evidence or argument, that NAEP results are a target and State results, where they differ from NAEP results, are inflated, distorted, ill-intentioned or suffering from low standards.

State performance standards are not simply statements about what students should know and be able to do. Because NCLB functions to focus attention on the proficiency cut score, State cut scores end up defining the portion of the student population for whom schools will be held most directly accountable. As states aim to improve the academic achievement of the disadvantaged under the auspices of Title I, it becomes reasonable to focus the accountability spotlight at the point where many disadvantaged students score. It is initially counterintuitive but nonetheless probable that a better way to achieve the goals of Title I is to set a lower proficiency standard.

In contrast, NAEP cut scores do not have high-stakes and do not function as policy tools. They are instead positioned as lofty, long-term goals. Even for this purpose, NAEP performance standards have attracted notice for their stringency. Richard Rothstein and his coauthors have argued persuasively that “NAEP cut scores for achievement levels are unreasonably high” (Rothstein, Jacobsen, & Wilder, 2006, p. 12). Regardless, with these perspectives, state performance standards can be seen to be justifiably lower than NAEP’s because they function to encourage commitment to educational equity for disadvantaged students.

In addition, the domains of Reading and Mathematics are vast. NAEP and State tests sample from these domains in a non-representative, non-identical fashion. As stakes increase on State results, results diverge from those of NAEP as one would expect from increased attention to a non-representative subset of the domain. Under this model, the divergence of results is not in and of itself evidence of inflation. Nor are higher percents of proficient students evidence of low standards. Performance standards may reasonably differ across subsets of the domain. It would strain credulity to argue that no state score inflation exists or that all state performance standards are set at reasonable levels. Our point is that the practice of defaulting to NAEP results as benchmarks is one that often lacks appropriate evidence, argument, and perspective.

The NAEP “audit” role has gained such widespread (though largely unexamined) acceptance that states are facing increasing pressure to align their content standards to NAEP frameworks. This move may be politically expedient, but NAEP frameworks were never intended to support or

match state priorities. To the extent of this mismatch, state policy makers would be doing their constituents a disservice.

The National Assessment Governing Board has largely tried to avoid high stakes for NAEP both in policy and in perception and has expressed cautions similar to ours in the past (National Assessment Governing Board, 2002). The desire to evaluate the effects of NCLB implementation across states has nonetheless dramatically increased attention to NAEP results. In spite of this, we do not believe that continued or increasing high stakes on NAEP results are a foregone conclusion. The movement toward State-NAEP comparisons has arisen, most fundamentally, from a lack of trust in state testing results. NAEP comparisons using the proficiency metric conveniently confirm discrepancies but, as we have repeatedly emphasized, do not in any way represent conclusive evidence that State results are suspect. As awareness of the fundamental differences between state tests and NAEP grows more widespread, we hope and anticipate that State-NAEP discrepancies will be used, not to confirm suspicions of invalid State results, but to begin deeper explorations into the differences between tests and testing practices. It is possible that these investigations will reveal inflated results or low standards, but State-NAEP discrepancies alone cannot support those conclusions.

References

- Braun, H. & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, 86(7):489-497. Retrieved in April, 2007, from http://www.pdkintl.org/kappan/k_v86/k0503br1.htm
- Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.). (in press). *Linking and Aligning Scores and Scales*. Springer.
- Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. Washington, DC: National Academy Press.
- Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4):237-261.
- Herman, J.L. & Webb, N.M. (Eds.). (2007). *Applied Measurement in Education*. 20(1) Lawrence Erlbaum Associates, Inc.
- Hill, R. (1998). *Using NAEP to Compare State Data—While It's Still Possible*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA. Dover, NH: Advanced Systems, Inc.
- Ho, A.D. & Haertel, E.H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples*. CSE Technical Report #665. University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA. Retrieved in April, 2007, from <http://www.cse.ucla.edu/products/reports/r665.pdf>

Holland, P.W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, 27(1):3–17.

Koretz, D. and Hamilton, L. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). American Council on Education and Praeger Publishers. Westport, Connecticut.

Linn, R.L. (2005). *Adjusting for Differences in Tests*. Paper prepared for Symposium on Use of School-Level Database. Washington, DC: The National Academies, Board on Testing and Assessment. Retrieved in April, 2007, from http://www7.nationalacademies.org/bota/School-Level%20Data_Robert%20Linn-Paper.pdf

McLaughlin, D. & Bandeira de Mello, V. (2005). *How to Compare NAEP and State Assessment Results*. Presented at the 35th Annual National Conference on Large-Scale Assessment. Retrieved in April, 2007, from http://38.112.57.50/Reports/LSAC_20050618.ppt

Musick, M. (1996). *Setting Education Standards High Enough*. Atlanta: Southern Regional Education Board.

National Assessment Governing Board. (2002). *Using the National Assessment of Educational Progress To Confirm State Test Results*. Retrieved in June, 2007, from http://www.nagb.org/pubs/color_document.pdf

Rothstein, R., Jacobsen, R., & Wilder, T. (2006). ‘*Proficiency for All*’ – *An Oxymoron*. Paper prepared for the symposium, “Examining America’s Commitment to Closing Achievement Gaps: NCLB and Its Alternatives.” Downloaded from http://www.epinet.org/webfeatures/viewpoints/rothstein_20061114.pdf in April, 2007.