(Over)-Interpreting Mappings of State Performance Standards onto the NAEP Scale

Andrew Ho *Edward Haertel*
University of Iowa *Stanford University*

In June, the National Center for Education Statistics (NCES) will release a report by Professor Henry Braun of Boston College and Dr. Jiahe Qian of Educational Testing Service (ETS). Among its many contributions, this report maps states' performance standards (e.g., proficient) onto the score scale of the National Assessment of Educational Progress (NAEP). At a glance, this mapping represents a ranking of states on the stringency of their performance standards. Its objective is to explain two provocative discrepancies in state-level test score reporting: 1) the often dramatic differences between states' percents proficient (state-state comparisons) and 2) the often dramatic differences between states' percents proficient and their corresponding NAEP percents proficient (state-NAEP comparisons).

Figure 1. Hypothetical Discrepancies in Percents Proficient Across States and Across Tests.

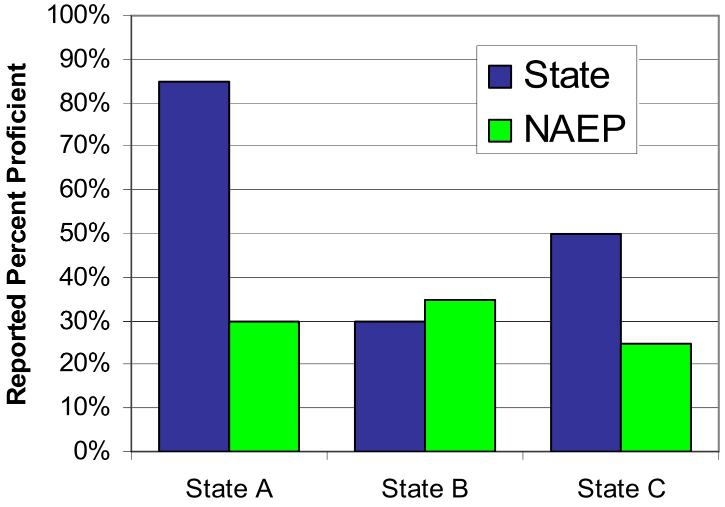


Figure 2. The Braun-Qian Explanation: State Performance Standards Mapped Onto the NAEP Scale, with NAEP Standards.

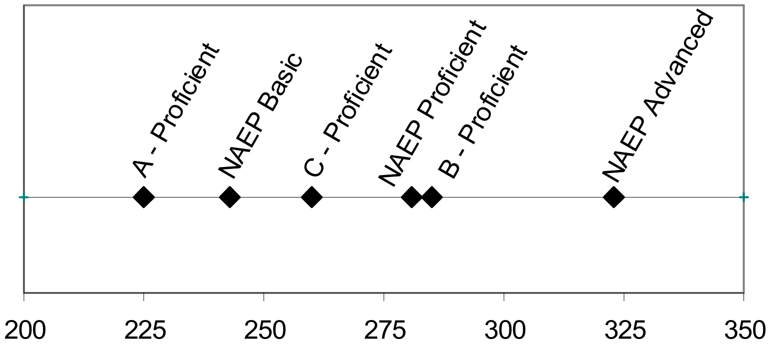


Figure 1 shows these discrepancies for three hypothetical states. Figure 2 shows how the Braun and Qian (hereafter abbreviated "BQ") method might explain these discrepancies. In Figure 1, state-state discrepancies are displayed in dark blue. The discrepancy is provocative because, at face value, it seems implausible that the 55 percentage point difference between States A and B is due solely to differences in student achievement. State-NAEP discrepancies are displayed as dark-blue/light-green pairs. At face value, it seems implausible that State A has 85 percent proficient students on the state test while only having 30 percent proficient students on NAEP. The BQ method generates mappings like Figure 2 that explain discrepancies between percents proficient as differences in the location of the proficiency cut score across states. State A's strikingly large percent of proficient students is thus explained by its low performance standard for proficiency.

This policy brief describes over-interpretations that may be encouraged by state-NAEP mappings like those in Figure 2 and offers some cautions concerning such interpretations. Braun and Qian have expressed similar caveats elsewhere—and we are confident that they will describe them in detail in their report—but their caveats are not expected to take the foreground or the exact form of the cautions expressed here. We have not yet seen the (currently embargoed) report.

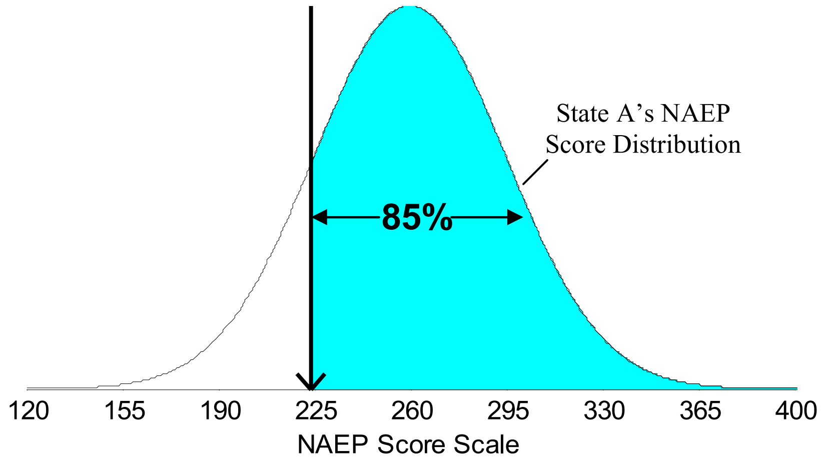
The primary argument of this policy brief is that interpretations of mappings like those in Figure 2 depend crucially on the often untested assumption that NAEP and state tests are equivalent. We demonstrate how seemingly straightforward interpretations like, "State A has a lower standard than State B," degenerate under plausible scenarios. Our intent is not to deny that differences in state standards exist. Rather, we caution that comparisons of standards become incoherent and misleading if the tests themselves do not function to measure the same achievement domain. We point out that the BQ mapping cannot address this necessarily prior concern on its own. A second, related policy brief reviews the topic of state-NAEP comparisons more generally and is entitled, "Apples to Apples? The Underlying Assumptions of State-NAEP Comparisons."

Comparisons of standards become incoherent and misleading if tests do not function to measure the same achievement domain.

The BQ Method and its Corollaries

The BQ method is a modification of a previous method developed by Drs. Don McLaughlin and Victor Bandeira de Mello at the American Institutes for Research (AIR). The original method and its modification produce very similar results; the differences between the methods will not be discussed here. The mapping addresses differences in state standards by making adjustments based on performance on a common test: NAEP. We describe the BQ method as a three-step process: Step 1) Within a state, take the full NAEP sample: the students from the schools and grades that took NAEP. Step 2) Take the percentage of these students who are proficient on the state test. Step 3) Find the NAEP cut score for that state that will set *the same percentage of students* as proficient. This score represents the state cut score mapped onto the NAEP scale. Figure 3 shows Step 3 of the BQ method for the hypothetical State A from Figures 1 and 2, resulting in a mapped standard of 225.

Figure 3. Step 3 of the BQ Method: A cut score of 225 passes the same percent of students as the state test.



Three illustrative corollaries follow from this approach. First, within a state, the larger the percent of proficient students in the Step 1 sample, the *lower* the NAEP mapped standard will be. This follows naturally from Figure 3, where one might visualize the state percentage of proficient students (85 percent) "pushing down" the mapped score, represented by the vertical arrow, from right to left. Across states, this means that, all else being equal, states that report greater percents proficient will have lower mapped standards. Second, if two states report the *same* percent of proficient students within the Step 1 sample, the state that performs *better* on NAEP will have the *higher* NAEP mapped standard. Across states, this means that, all else being equal, higher scoring NAEP states will have higher mapped standards. Third and finally, if a state with a low mapped standard wanted its mapped standard to match NAEP's, it would have to raise its state cut score (pass fewer students) until the Step 1 sample's percent of proficient students fell to match NAEP's.

Considering these corollaries together, the BQ method can be seen as a standardization of state performance standards based on NAEP performance, including negative adjustments or "penalties" for reporting high percents of proficient students without commensurately high NAEP performance.

The mapping essentially penalizes state performance standards for reporting high percents of proficient students without commensurately high NAEP performance.

Caveats for Cross-State Comparisons

As McLaughlin and Bandeira de Mello have noted, the argument for penalizing high state percents proficient based on NAEP performance relies logically on the relationship between the state test and NAEP. If the achievement represented by a state's "proficient" standard cannot be meaningfully expressed in terms of NAEP *content* standards, the mapping loses its justification. The same issue arises in comparing states' "proficient" standards to one another: If we try to map one state's definition of proficiency onto a scale for a *different* set of achievements, we might reasonably conclude that no valid mapping exists. As an extreme but illustrative example, consider mapping state *football* standards onto the NAEP scale. The Step 1 sample has a percentage of students who are considered proficient at football. That percentage (Step 2) can be mapped on to the NAEP scale (Step 3) and compared across states. However, the corollaries quickly lose their rationale. Why should performance standards for football be penalized if high percents of football proficiency do not correspond with high *NAEP* performance?

This analogy is exaggerated but emphasizes the fact that the validity of the BQ mapping method depends on a necessarily prior question: Can each state's performance standards be meaningfully expressed in terms of NAEP curriculum frameworks? Indeed, the validity of the full mapping depends on the meaningful expression of all performance standards in terms of all other state tests' content standards, something that would be intractable if the tests were as dissimilar as football, baseball, and tennis, for example. Table 1 shows sports analogies for the logic of state-NAEP standard mapping. Some of them are deliberately absurd to highlight the point, others may be more apt.

Table 1. More and Less Apt Sports Analogies
for State-NAEP Standard Mapping

State - NAEP Comparisons	Penalize State performance standards for reporting high percents of proficient students without commensurately high NAEP performance.
Football - NAEP Comparisons	Penalize State football standards for reporting high percents of football proficiency without commensurately high NAEP performance
Golf - Miniature Golf Comparisons	Penalize State Golf standards for reporting high percents of golf proficiency without commensurately high miniature golf performance
Hard Court - Clay Court Tennis Comparisons	Penalize State tennis standards for reporting high percents of hard court tennis proficiency without commensurately high clay court tennis proficiency.

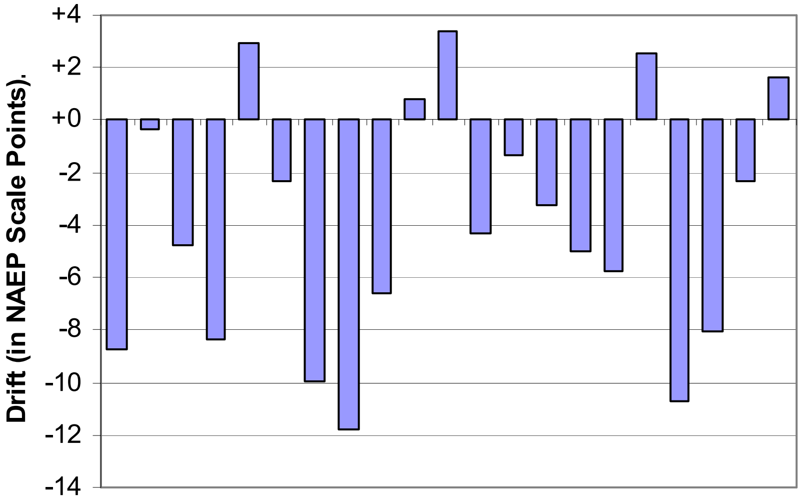
Regardless, the BQ mapping procedure cannot discern or confirm the relationships on its own. In fact, Braun, Qian, McLaughlin, and Bandeira de Mello have access to evidence of the correspondence between state tests and NAEP in the form of school-level correlations. However, it is revealing that the BQ mapping can be shown to be invariant to these correlations.

In sum, the act of comparing percents proficient across tests takes for granted the equivalence of the score scales on which different “proficiency” cut scores are set. In practice, this equivalence is rarely explicitly defended. If the abstract notion of proficiency itself is used to support a common scale without considering the equivalence of the domain, this may be accurately described as the logical fallacy, “begging the question.” In short, comparing proficiency standards begs the question, “Proficient at what?”

The Stability of the BQ Mapping over Time

A growing body of literature has consistently found that state and NAEP trends are often dissimilar and occasionally disordinal. As Braun and Qian have noted in previous presentations, these “trend discrepancies” have important implications for the stability of the BQ mapping over time. If the state percent of proficient students rises without a corresponding NAEP increase, Step 3 of the BQ method will lead to a lower mapped standard (see Figure 3). The mapped standard will continue to decline as long as the state trend exceeds a function of the NAEP trend. Given that other states have different state-NAEP trend discrepancies, the ordering of states by their mapped standards will change over time.

Figure 4. A Rough Estimate of the Degree of Mapped Standard Drift for 21 States Due to Trend Discrepancies



Estimates of the degree of the drift of mapped NAEP scores are shown in Figure 4. These are rough estimates based on real trend discrepancy findings for 21 anonymous states with appropriate grade 8 reading test results for both state and NAEP tests from 2003 to 2005. The range of possible drift is as positive as almost 4 points (a state whose NAEP trend exceeds its state trend) and as negative as almost 12 points (a state whose state trend exceeds its NAEP trend) on the NAEP score scale. With NAEP standard deviations of 35 for grade 8 reading, this drift can appear substantial. These estimates are very rough given that matched school-level data were not available for Step 1 of the BQ method, but the degree of performance standard drift is not expected to change with a more rigorous application of the method.

If a state’s mapped standard drifts over time, in what sense has its performance standard actually changed? The state test score scale and cut scores have not changed, nor have the score scale and cut scores for NAEP. The changes in this mapping over time are an undesirable property of the BQ method for which there is no obvious solution. It arises naturally from discrepant trends, which in turn may arise from non-identical test content, changing teaching practices, differential changes in student motivation, and many other possible factors and their interactions. Drift underscores the take-home point from the previous section: Comparing performance standards assumes similar tests. When cross-test comparisons show evidence of test non-equivalence, as trend discrepancies do, the latent flaws in interpretations of mappings become manifest.

One might be tempted to interpret a declining mapped standard simply as a declining standard. This is terribly misleading; the cut score has not actually changed. The core issue in declining mapped standards is that achievement as measured by the state test has risen faster than achievement as measured by NAEP. Considering the accountability structure under No Child Left Behind, this kind of discrepancy—and therefore drift in mappings—could be framed as precisely what one would expect. Once one appreciates the scope of the domains of reading and mathematics and the degree to which test content frameworks do not overlap, marginally discrepant trends on tests seem less controversial.

Conclusions

In summary, Braun and Qian have developed an attractive and technically sound mapping whose most basic interpretations mask strong assumptions about the functional similarity of reading and mathematics assessments both across states and between states and NAEP. We anticipate that these assumptions will be largely ignored by media reports that will simply laud states with high mapped standards and chide states with low mapped standards. NAEP will receive credit for having such high standards. Continued attention to these judgments could result in political pressure to "race for the bottom," where states try to artificially pass fewer and fewer students in an attempt to seem rigorous that is largely devoid of substantive considerations about the domain.

We encourage framing the BQ mapping as exactly what it is: a system for adjusting state proficiency results where they are not reflected by corresponding NAEP proficiency. For the group of states with NAEP-like frameworks and NAEP-like tests, these comparisons are genuinely meaningful and should spark useful discussions about definitions of proficiency. For the group of states with evident differences between state and NAEP frameworks, implementation, and stakes, state mappings encourage comparisons that cannot be defended.