Research Report

Measurement Issues in State Achievement Comparisons

Rebecca Zwick

July 2010

ETS RR-10-19

ETS®

Listening. Learning. Leading.®

Research Report

Measurement Issues in State Achievement Comparisons

Rebecca Zwick

July 2010

ETS RR-10-19

ETS®

Listening. Learning. Leading.®

Measurement Issues in State Achievement Comparisons

Rebecca Zwick
ETS, Princeton, New Jersey

July 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Shelby J. Haberman

Technical Reviewers: Mary Pitoniak and Randy Bennett

Copyright © 2010 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

SAT is a registered trademark of the College Board.

Abstract

Political perspectives on the advisability of state achievement comparisons have changed substantially over the last half-century. As Selden (2004) noted regarding the National Assessment of Educational Progress (NAEP), "[w]hen it began in the 1960s, NAEP specifically was designed *not* to provide comparisons among states. This was done for political reasons, so that the assessment program would be palatable to the educational community and to the states" (p. 195). During the No Child Left Behind Era, state comparisons were complicated by the use of tests and proficiency standards that differed across states. Because of today's Common Core Standards Initiative, state comparisons are once again at the forefront of educational policy discussions. The initiative brings with it both significant opportunities and substantial psychometric challenges.

Key words: achievement, Common Core Standards, NAEP, No Child Left Behind, proficiency, Race to the Top, Standards, State assessment

Acknowledgments

This paper is an updated version of a presentation at the National Research Council workshop, Best Practices for State Assessment Systems, Washington, DC, December 11, 2009. The paper benefitted from ETS reviews by Randy Bennett and Mary Pitoniak. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the Division of the Behavioral and Social Sciences and Education or of the National Academy of Sciences.

The December 2009 workshop of the National Research Council, Best Practices for State Assessment Systems, focused on the Common Core State Standards Initiative (CCSI), a joint effort of the National Governors Association and the Council of Chief State School Officers in partnership with Achieve, ACT, and the College Board. This effort, initiated in late 2008, has energized the public conversation about the possibility of comparing states' academic performance on a common assessment based on a shared set of educational standards.

Political perspectives on the advisability of state achievement comparisons have changed substantially over the last half-century. An examination of the history of the National Assessment of Educational Progress (NAEP) reveals a particularly dramatic shift between the 1960s and the 1980s. As Selden (2004) noted, "[w]hen it began in the 1960s, NAEP specifically was designed *not* to provide comparisons among states. This was done for political reasons, so that the assessment program would be palatable to the educational community and to the states" (p. 195). Messick, Beaton, and Lord (1983, p. 16) noted that in the 1960s, the U.S. government was viewed as intruding on the educational realm because of such actions as federal enforcement of school desegregation in the wake of the Civil Rights Act of 1964 and Congressional passage of the Elementary and Secondary Education Act in 1965.

The political climate had changed vastly by the 1980s. Although the Reagan administration deemphasized federal education programs, even threatening to eliminate the Department of Education, it nevertheless promoted the idea of academic excellence through competition. In 1983, the National Commission on Excellence in Education (1983) issued its landmark report, *A Nation at Risk*, which declaimed, "If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war" (para. 2). President Ronald Reagan's 1984 *State of the Union* address called for comparing achievement for states and schools: "Without standards and competition there can be no champions, no records broken, no excellence. . . ." The same year, the Council of Chief State School Officers established its State Education Assessment Center and initiated discussions on procedures for making state achievement comparisons (Ferrara & Thornton, 1988).

The year 1984 also marked the beginning of the infamous wall charts of the U.S. Department of Education. These charts, which continued to appear through 1991, compared states in terms of educational resources and educational performance, including average SAT®

and ACT test scores. These comparisons were disturbing for several reasons. First, the tests were college admission tests, rather than assessments aligned with K-12 instruction. Second, the students taking the tests were primarily those who intended to go to college, rather than a representative sample of high school students in general. And third, the states differed widely in terms of the percentages of students taking the tests. Although some have described the wall charts as "a bad idea done badly" (Ferrara & Thornton, 1988, p. 200), a study group conducting a review of NAEP offered a more benign perspective. According to the Alexander-James committee, chaired by the governor of Tennessee, Lamar Alexander, and H. Thomas James, "[t]he problem with [the Secretary of Education's] wall chart is that its sources of performance ... information are wretchedly inadequate. We know why … there are simply no other comparable data available" (Alexander & James, 1987, p. 6).

Regarding the administration of NAEP, the committee stated, "The single most important change recommended by the Study Group is that [NAEP] collect representative data on achievement in each of the fifty states and the District of Columbia" (Alexander & James, 1987, p. 11). The following year, new NAEP legislation authorized the initiation of a trial state assessment to be conducted in 1990. Thirty-seven states participated, administering a math assessment to eighth graders. By 1996, NAEP state assessments had become an official part of the program and were no longer considered "trial." Even with its addition of a state component, however, NAEP was not fully responsive to Reagan's call to action in his *State of the Union* because it did not provide scores for individual students and, at that time, was typically viewed as low-stakes. In his own *State of the Union* address 13 years later, President Bill Clinton proposed that states adopt, within two years, national tests of student achievement in reading and math, a program subsequently labeled the Voluntary National Test. After much congressional wrangling, the test was called off for lack of funding in 1999.

Policies involving state assessments continued to evolve in the early 21st century. In a significant move in the direction of state comparisons, the No Child Left Behind Act of 2001 required states receiving Title I funds to participate in state-level NAEP assessments in reading and mathematics at grades 4 and 8 every two years. As part of its test-based accountability and sanctions system for states, NCLB also resulted in a second type of state comparison: States' proficiency rates were compared using the results of the states' own assessment programs, despite differences among the states in terms of standards, tests, and definitions of proficiency.

Studies comparing state proficiency cut-points to those used by NAEP have revealed that states' definitions of proficiency vary widely and tend to be less stringent than those used in NAEP (see National Research Council, 2008, for a review of several such studies). For 2005 results on grade 8 reading, Barton (2009, p. 18) compared proficiency rates (the percentage of students considered proficient or higher) determined by state assessments to those obtained by NAEP. Comparisons were available for 33 states and the District of Columbia. While NAEP's proficiency rates ranged from 12% to 38%, the states' rates ranged from 30% to 88%. Nor is it the case that states with similar NAEP proficiency rates had similar state-determined rates. For example, Florida and Georgia both had NAEP proficiency rates of 25%. However, based on their own states' results, Florida's proficiency rate was 43%, while Georgia's was 83%.

As the recent National Research Council (NRC; 2008) report, *Common Standards for K-12 Education? Considering the Evidence*, noted, "The variation in student performance has caused many to wonder whether the logical next step for a nation committed to improving achievement for all students is to move toward common standards" (p. 2).

Today's CCSI creates both opportunities and challenges for state achievement comparisons. At the NRC workshop, Best Practices for State Assessment Systems, these opportunities and challenges were discussed by Hambleton (2009) and Wise (2009), among other presenters. In the following sections, these opportunities and challenges are reviewed with respect to their implications for assessment. In the subsequent discussion, it is assumed that states that form a consortium subscribing to a set of common standards will wish to compare achievement across the participating states using tests that produce individual student scores. It is further assumed that these participating states will wish to adopt a common definition of proficiency based on these scores.

Common Core State Standards Initiative (CCSI): Opportunities for Assessment

One benefit of the CCSI is a likely boost in the funding that is available for assessment and assessment-related research. By grouping together in consortia, such as the New England Common Assessment Program, states can take advantage of economies of scale in test development, as noted by Wise (2009). The degree to which a state has committed to common standards is being evaluated in the competition for federal Race to the Top funds. In the Phase 1 competition, for which awards were made in 2010, the Federal Register announcement indicated

that reviewers would consider the "extent to which the State has demonstrated commitment to improving the quality of its standards by participating in a consortium of States that is working toward jointly developing and adopting … a common set of K–12 standards … and the extent to which this consortium includes a significant number of States" (Federal Acquisition Regulations for Department of Education Race to the Top Fund, 2009, p. 37804).

Ideally, increased funding for assessment could help to produce state tests that are better aligned with standards, contain richer and more innovative questions, and are more reliable (since it is more feasible to develop an adequate number of items). Reporting could be made more timely, and diagnostic information and instructionally relevant feedback (i.e., the formative aspects of assessment) could be enhanced (Wise, 2009). The recent NRC (2008) report on common standards noted that "few states systematically provide for extensive formative assessments that teachers [can] use to tailor instruction to individual students' needs … States could much more easily take advantage of one another's knowledge and experience, and avoid duplication of effort, if they were applying consistent frameworks" (p. 23).

In addition, a boost in funding could promote better research on tests, which could lead to further improvements (Wise, 2009). This research could include cognitive analyses of test content, piloting of items, studies of validity, investigations of fairness of tests for all ethnic, language, socioeconomic, and gender groups (e.g. studies of differential validity and differential item functioning), and studies of testing accommodations.

Finally, a frequently cited advantage of common standards and assessments is the political cover it provides. When standards and assessments are shared, it is less likely that any one state will reduce its expectations of student performance in order to achieve nominally higher proficiency rates.

Common Core State Standards Initiative (CCSI): Challenges for Assessment

Attempting to create an assessment that is common across multiple states is a difficult and complex enterprise: How can we ensure that state achievement comparisons are valid, fair to the participating states, and ultimately, useful for improving teaching and learning? This section discusses challenges involving measurement and interpretation issues.

A key question regarding the success of state achievement comparisons is whether an assessment program can be developed that allows states within the same consortium to claim that

their results are "on the same scale." Although achieving this goal would not necessarily require an identical test to be administered in all participating states, a substantial number of common items would need to be developed in each grade and subject area in order to link the overlapping tests. This task is by no means trivial or straightforward, especially since the likely overlap among state standards is far from clear, as detailed below.

The Race to the Top guidelines state that a "[c]ommon set of K–12 standards means a set of content standards that define what students must know and be able to do, and that are identical across all States in a consortium. Notwithstanding this, a State may supplement the common standards with additional standards, provided that the additional standards do not exceed 15 percent of the State's total standards for that content area" (Federal Acquisition Regulations for Department of Education Race to the Top Fund, 2009, p. 37811). This guideline, despite its mention of identical standards, does not actually ensure a high degree of overlap in standards among all states in a consortium, even under the assumption that these states wish to be eligible for Race to the Top Funds. This can be illustrated by the following example: For the sake of simplicity, suppose that we quantify standards in a particular subject area and grade simply by enumerating them (clearly an inadequate method since some may be much broader and more important than others). Assume that the 10 states in a particular consortium share 85 standards in common. Each state also has 15 unique (state-specific) standards, thus complying with the 15% rule. Then out of the total of $85 + 10(15) = 235$ standards, only $85/235(100) = 36.2\%$ are shared among the 10 states. And the 15% rule says nothing about how much of the assessment itself must be shared among states within a consortium.

An obvious impediment to the development of common standards and common test items is the difficulty of reaching consensus across states. As Ferrara and Thornton (1988) noted in connection with NAEP's early state assessment efforts, debates about assessment content are likely to be "fired with a new intensity because of impending interstate comparisons" (p. 208). If the consensus process is reduced to a tabulation of objectives that are already common across participating states, the effect will be to narrow the focus of the test content as well as the cognitive complexity of the items, thus compromising the CCSI's potential for improving teaching and learning.

Even if a large number of common items can be created, it is possible that the inclusion of state-specific items in the assessment could affect performance on common items. Such

context effects have been widely documented in the educational measurement literature (Brennan, 1992). Even changing the position of an item can have a substantial effect on performance (e.g., see Hill, 2008, pp. 7-8), as was the case in the NAEP reading anomaly (Zwick, 1991) in which the performance of 9- and 17-year-olds showed an unexpected drop between 1984 and 1986. Another peril related to state-specific items is that they may measure somewhat different skill and content dimensions from the common core material, impairing the ability to successfully link the states' overlapping tests and to report the results for these states on the same scale (Hambleton, 2009).

An additional measurement challenge involves the degree to which states are willing and able to adhere to common testing policies and procedures. For example, in order for results to be comparable across states, agreement is needed on the timing and conditions of test administration (including whether the test is computerized), the rules for granting testing accommodations, the policies determining which students are to be considered exempt from testing or reporting requirements, and the types of test preparation that are considered permissible.

Even if states adhere strictly to such protocols, comparing results across states has the potential to be misleading. In 1987, the National Academy of Education (NAE) expressed some misgivings about the recommendation of the Alexander-James commission that NAEP conduct state-by-state comparisons. The NAE stated, "We are concerned about the emphasis [in the report] on state-by-state comparisons of average test scores. Many factors influence the relative rankings of states ... Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts" (Glaser, 1987, p. 59).

Indeed, it is true that interpreting state achievement differences is complex, and this complexity would be diminished, but certainly not eliminated, under the CCSI. Because curriculum and instruction would still be likely to differ under common standards, opportunity to learn would be expected to vary as well. Of course, states also differ in terms of demographic characteristics, per-pupil expenditures, teacher licensing and training, class size, student retention policies, and countless other factors.

In similar contexts, statistical adjustments are often proposed to achieve a more fair ranking of states. However, adjustment models can themselves be controversial. And in past examples, such as attempts to statistically adjust states' average ACT or SAT scores to correct for differing participation rates, success has been limited. In fact, competing adjustment methods

based on seemingly sensible assumptions have been found to produce very different rankings of the 50 states (see Wainer, 1989).

Given the daunting list of challenges in conducting a common assessment program and interpreting the results, how can meaningful achievement comparisons be performed?

Recommendations for Meaningful Achievement Comparisons Across States

Although the task of developing a common assessment for use in comparing states is unquestionably a difficult one, state consortia can take steps to increase the likelihood of success. The following sections address recommendations concerning the test instrument, the testing policies, and the interpretation of test results, respectively.

The Test Instrument

To ensure that reliable scores can be reported and that linkages among states within a consortium can be achieved, it is essential that an adequate pool of items be developed within each grade and reporting area. A reporting area may be narrower than the nominal subject of the assessment, such as mathematics: If subscores for certain skills or topics are desired, a sufficient number of items must be developed within each skill or topic area. Because a sound item development process is contingent on plans for score reporting, it is important that at least a tentative decision about scoring and reporting models be made as soon as is practicable.

Maximizing comparability of scores across states requires not only a sufficient number of common items, but uniformity of test format and instructions. Psychometric research has repeatedly shown that even subtle changes in test layout can affect performance. Any items that are not a part of the common core of the assessment should appear after the common items, preferably in a separate section, in order to minimize item context effects.

Finally, any funds that become available because of economies of scale or as a result of new federal education efforts should be fully exploited to conduct research that could potentially improve the quality of the assessment.

Testing Policies

Maximizing comparability across states requires that testing policies, as well as the instrument itself, be tightly controlled. States will need to implement rigorous and detailed protocols in a number of areas. Test administration, including the setting and dates of

administration, length and number of sessions, and administration mode should be specified so that consistency can be maintained across states within a consortium. In particular, computerized administration should not be assumed equivalent to paper-and-pencil administration without rigorous supporting research (see Hambleton, 2009). Another significant aspect of testing policy concerns the rules that govern exclusions and accommodations for English language learners and students with disabilities. Which students can legitimately be excluded from testing or reporting requirements? Which students are eligible for accommodations? Which accommodations are appropriate? These complex issues continue to be troublesome in many large-scale testing programs (Koenig & Bachman, 2004; Shakrani & Roeber, 2009), but clearly need to be resolved in order to make meaningful comparisons across states.

States within a consortium will also need to work together to delineate what test preparation and motivational activities are permissible, what scoring procedures are to be used for items that require human raters, and what test security procedures are to be implemented.

Interpretation

To increase the likelihood that state comparisons will be meaningful and useful, consortia can take two types of actions. First, they can provide as much context as possible for these comparisons by documenting differences across states in demographics, resources, instruction, and educational policies. Protocols for defining and reporting these contextual variables will be needed to ensure that data on all relevant characteristics are collected in a rigorous and comparable fashion. (For example, one state's per-pupil expenditure should not be compared to another state's per-pupil *instructional* expenditure.) A second step that consortia can take to discourage incorrect or simplistic interpretation of academic performance differences across states is to invest resources in improving the capacity of educators at all levels to interpret test results. A host of studies in the last decade have revealed large deficits in teachers' and administrators' skills in comprehending and using test results (see Zwick et al., 2008), despite the large increase in the amount of testing data available to educators. This finding is not entirely surprising, at least in the case of teachers, given that a recent survey showed that no state required successful completion of a course in assessment for teacher certification (Stiggins & Herrick, 2007).

Summary

In summary, perspectives on cross-state achievement comparisons have changed substantially since the inception of NAEP almost 50 years ago. At that time, the program's development of "a sampling plan insuring that accurate results could not readily be reported at the state or district level" was "brilliantly responsive to the political constraints of the time" (Messick et al., 1983, p. 11). Today, the fact that 48 states have signed onto the CCSI suggests a willingness to at least consider state achievement comparisons. These comparisons would likely be made within consortia of states subscribing to a common core of standards. For a state, joining a collaborative venture of this sort could bring with it substantial opportunities for increasing assessment resources. However, the enterprise also involves significant challenges in the areas of measurement and interpretation. Some steps that a consortium can take to improve the chances of developing meaningful and fair state achievement comparisons that are useful for instruction are the following:

- Make the test instrument and testing policies as similar as possible for states within a consortium
- Provide ample context for the achievement comparisons by including information on state demographics, resources, instructional practices, teacher training, and other relevant variables
- Offer high-quality instruction to all teachers, school administrators, and state education officials in the interpretation and use of test scores, including the underlying measurement and statistics principles
- Finally, as recommended by Hambleton (2009), foster collaboration among psychometricians, educators, and policy makers, who, despite their differences in perspective, are likely to be pursuing the same ultimate goal—the improvement of teaching and learning in the nation's classrooms

References

Alexander, L., & James, H. T. (1987). *The nation's report card*. Cambridge, MA: National Academy of Education.

Barton, P. E. (2009). *National education standards: Getting beneath the surface* (ETS Policy Information Center Report). Princeton, NJ: ETS.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5,* 225–264.

Ferrara, S. F., & Thornton, S. J. (1988). Using NAEP for interstate comparisons: The beginnings of a "national achievement test" and "national curriculum." *Educational Evaluation and Policy Analysis, 10*, 200–211.

Federal Acquisition Regulations for Department of Education Race to the Top Fund, 74 Fed. Reg. 37804 (proposed July 29, 2009).

Glaser, R. (1987). A review of the report by a committee of the National Academy of Education. In L. Alexander & H. T. James (Eds.), *The nation's report card* (pp. 43–61). Cambridge, MA: National Academy of Education.

Hambleton, R. K. (2009, December 11). *Using common standards to enable cross-state comparisons.* Presented at the National Research Council workshop, Best Practices for State Assessment Systems, Washington, DC.

Hill, R. (2008, June 17). *Using p-value statistics to determine the believability of equating results.* Paper presented at the National Conference on Student Assessment, Orlando, FL.

Koenig, J. A., & Bachman, L. F. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment.* Washington, DC: The National Academies Press.

Messick, S., Beaton, A., & Lord, F. (1983). *A new design for a new era.* Princeton, NJ: National Assessment of Educational Progress and ETS.

National Commission on Excellence in Education. (1983). *A nation at risk.* Retrieved from the U.S. Department of Education website: http://www.ed.gov/pubs/NatAtRisk/risk.html

National Research Council. (2008). *Common standards for K-12 education? Considering the evidence: Summary of a workshop series*. Washington, DC: The National Academies Press.

Reagan, R. (1984). *State of the union address.* Retrieved January 13, 2010, from the Website of The American Presidency Project, http://www.presidency.ucsb.edu/ws/index.php?pid=40205

Selden, R. (2004). Making NAEP state-by-state. In L. V. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 195–199). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Shakrani, S. M., & Roeber, E. (2009). *Suggested model rules for uniform national criteria for NAEP testing in national and state samples.* Washington, DC: National Assessment Governing Board. Available for download at http://www.nagb.org/publications/shakrani-roeber-uniform-naep.doc

Stiggins, R., & Herrick, M. (2007). *A status report on teacher preparation in classroom assessment*. Unpublished manuscript.

Wainer, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics, 14* (2), 121–140.

Wise, L. L. (2009, December 11). *How common standards might support improved state assessments.* Presented at the National Research Council workshop, Best Practices for State Assessment Systems, Washington, DC.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10,* 10–16.

Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice, 27*, 14–27.