Why Continue An Old Assessment?

A paper on the NAEP Long-Term Trend Assessments prepared for
the National Assessment Governing Board
Jack Jennings
February 13, 2017

In this age of student testing mania, a strong justification should be required to initiate or continue any assessment. Edward Haertel's excellent paper for the National Assessment Governing Board (Governing Board) makes such a case for the Long-Term Trend (LTT). My paper supplements Dr. Haertel's, emphasizing certain of his points and adding others from my experiences with the National Assessment of Educational Progress (NAEP).

I am not an expert in psychometrics; rather, my career has been in the policy arena. While working for the U.S. House of Representatives from 1967 to 1994, I helped write education laws, including those affecting NAEP. From 1995 to 2012, I monitored NAEP as part of my responsibilities heading the Center on Education Policy. Also relevant are my nine years of service on the Board of Trustees of the Educational Testing Service, which has the principal contract dealing with NAEP. These experiences have shaped my views of NAEP.

My position is that the LTT is irreplaceable in understanding students' academic achievement in the United States. The Governing Board should thus reverse its decisions resulting in a 12-year hiatus in the administration of that assessment. Such a long break in data collection undermines the usefulness of the LTT.

The arguments to support the LTT include the Governing Board's legal obligation to maintain it, as well as prudence in retaining the only continuous measurement since the early 1970s of what students *actually* know and are able to do, while what students *should* know and be able to do has been measured from the early 1990s by the main NAEP.[i] Also noteworthy is the LTT's relevance in policy-making and the additional opportunities it affords to implement the Governing Board's new strategic vision.

I commend the Governing Board for organizing this review of its decision regarding the Long-Term Trend. The Governing Board is setting an excellent example for prudent policy-making by inviting a diverse set of people to critique its actions.

The Legal Obligation to Continue the LTT

A basic duty.

"The Commissioner for Education Statistics shall, with the advice of the Assessment Board… continue to conduct the trend assessment of academic achievement at ages 9, 13, and 17 for the purpose of maintaining data on long-term trends in reading and mathematics." The Governing

Board is therefore legally bound to continue the LTT, but not required to administer the assessment on any particular schedule.[ii]

The decisions resulting in a 12-year gap between administrations of LTT, however, indirectly frustrates that clear legal obligation to "continue to conduct the trend assessment..." The letter of the law is being carried out, but not its spirit.

When the Governing Board made these decisions, scarcity of funds was cited. Money, however, was found for a Technology and Engineering Literacy Assessment and other projects.

Over the years, the Governing Board's principal responsibility has been to oversee the assessment of the reading and mathematics achievement of American students through the LTT and the main NAEP. Other projects should not be funded unless that basic responsibility has been carried out.

Reasons for this particular responsibility.

The Governing Board's obligation to continue the LTT was imposed primarily because of the value of such long-term trends. An additional reason was concern about the development and use of the achievement levels in the main NAEP. A short review of NAEP's history will help to explain the latter issue.

In the 1988 education amendments, the Governing Board was created and charged with identifying appropriate achievement goals. The Governing Board established three: *Basic*, *Proficient*, and *Advanced*. An independent evaluation of the "trial assessments," as the new tests were known, was to be undertaken by a nationally recognized organization to assess the feasibility and validity of the assessments and the fairness and accuracy of the resulting data.[iii]

In 1990, the Governing Board hired a team of evaluators to study the process of comparing student performance using the three levels. This group was fired a year later after its draft report concluded that the process "must be viewed as insufficiently tested and validated, politically dominated, and of questionable credibility."[iv] The Governing Board disputed the assertion that the dismissal was due to these conclusions.

In 1993, the National Academy of Education (NAE) and the U.S. General Accounting Office reviewed the student performance standards concluding that they did not meet technical expectations and should not be used as the primary means for reporting NAEP.[v] The following year, a law was enacted requiring the Governing Board's performance levels to be labelled "on a developmental basis," until such time as an independent review by an organization such as the NAE or the National Academy of Sciences (NAS) found them to be reasonable, valid, and informative to the public.[vi]

In 1998, another congressionally mandated evaluation by the NAS concluded that NAEP's procedures for setting cut-scores was fundamentally flawed because it rested on "informal judgment" rather than a "highly objective process," and thus produced some unreasonable results. Highly critical of NAEP's achievement levels, the Academy urged caution in their use.[vii]

In 2001, the *No Child Left Behind Act* (NCLB), while requiring states to participate in NAEP, continued the achievement levels on a "trial basis." The law further required that the results from

the newly mandated state tests be reported using the same format as NAEP's three achievement levels. An important point, though, is that *NCLB* allowed states to have their own definitions of the level of achievement for each level.[viii] In other words, the labels were the same, but the definitions of achievement could—and did—differ by state from the national assessment.

In 2005, U.S. Secretary of Education Margaret Spellings urged reporters to compare state proficiency levels under *NCLB* to NAEP's *Basic* achievement level. Her statement was in response to the confusion resulting from the use of the same three titles for measuring student achievement in NAEP and in the *NCLB*-required state tests. Some states defined proficient on their tests to be higher than what was considered *Proficient* on NAEP, but most states' definitions were less demanding, resulting in a muddle. Spellings's view implied that NAEP's *Proficient* level was ambitious, more than what would be expected of most students.[ix]

In 2008, the Center for Public Education of the National School Boards Association sought to clarify the issue using Governing Board and U.S. Department of Education resources. The proficient level defined by states for their tests meant meeting grade-level expectations. In contrast, NAEP's *Proficient* level was an *aspirational goal* for American students, not grade-level achievement.[x]

These explanations were helpful; but when NAEP results were released, the news media continued to emphasize the percentages of students meeting the *Proficient* level, not those meeting the *Basic* level. Since *NCLB* had set the national goal of all students being proficient by 2012, proficiency seemed the objective—rather than meeting a basic level of academic performance.

In 2016, Campbell Brown, a former CNN anchor starting her own education reform group, said that two out of three eighth graders could not read or do mathematics at grade level. When challenged by Tom Loveless of the Brookings Institution about the accuracy of that statement, her response identified NAEP as her source. Loveless countered that *Proficient* in NAEP meant mastery over challenging subject matter, not doing mathematics or reading at grade level. She retorted: "But any reasonable person or parent can rightly assume that if their child is not reading at grade level, then their child is not proficient."[xi]

Also last year, another NAS report was issued on the NAEP achievement levels for reading and mathematics.[xii] The U.S. Department of Education had commissioned this review because these levels, more than two decades after their creation, were still labeled "trial." That designation resulted from the lack of an independent evaluation determining these levels to be "reasonable, reliable, valid, and informative to the public," as required by the 1994 law.

The Academy concluded that some aspects of the original process of determining the levels were positive. But since then, problems persisted with the levels' validity, lack of interpretive guidance for the results, and misalignment of the assessments. Hope was offered, though, that if certain steps were taken, such as better alignment "among the frameworks, the item pools, the achievement-level descriptors, and the cut scores," the levels could be considered as meeting the legislated standards.

This short history shows two major tensions. First, student achievement levels were adopted as a means of making more understandable NAEP student achievement data. Yet, 25 years after their creation, it is still a challenge to convey to the public and to the news media what NAEP has found. Second, independent expert criticisms during the same quarter-century have persistently found that the levels are lacking in certain key respects. Thus, they are still on "trial."

In the late 1980s, when these achievements levels were authorized by law, Congress realized that this process was something new and would be difficult to get right. The LTT, therefore, was continued as a *safeguard* so that there would be some way to measure achievement while the new process was fought over and developed. Since the National Academy of Sciences in last year's report still sees problems with the levels, we have not yet reached the end-stage of that development. Thus, the safeguard of the LTT is still needed.

Also important to note in continuing the LTT is the difference in purpose between the LTT and the main NAEP as it is usually reported in the news. As the latest NAS report states:

> Originally, NAEP was designed to measure and report what U.S. students *actually* know and are able to do. However, the achievement levels were designed to lay out what U.S. students *should* know and be able to do. That is, the adoption of achievement levels added an extra layer of reporting to reflect the nation's aspirations for students.

It is true that since the early 1990s the main NAEP has also reported on what students have actually learned. But, the news media concentrate on the achievement levels in their reporting of NAEP results. In fact, the main NAEP is defined by these achievement levels which were created to be aspirational or something that would motivate students to strive to do better. That purpose is quite different from reporting on what students have actually attained without regard to what they should have attained.

Relevance

Congress had its reasons for imposing the legal obligation on the Governing Board to continue the LTT. A major benefit from that decision is LTT's relevance in making policy.

The last few years have been a time of frustration for many state education leaders. Bush's *NCLB* and Obama's Race to the Top were seen as federal mandates, disrespectful of state and local control of education.

Recently, the National Conference of State Legislators (NCSL) decided to take matters into its own hands. On a bipartisan basis, leading legislators set out to determine how best the states themselves could improve elementary and secondary education. They began by studying NAEP's LTT and several international studies. From this review, the group concluded that American students are struggling to meet relatively low expectations.

No Time to Lose,[xiii] NCSL's first report from this project, prominently displays at the beginning of its analysis a table of LTT's data on student achievement. The report proceeds to recommend major systemic changes in American schools.

This attention to the LTT demonstrates its continuing usefulness to policymakers. Very appropriately, the Governing Board describes that assessment "as the largest, nationally representative, continuing evaluation of the condition of education in the United States." Other long-term student data sets have major limitations, for example, SAT and ACT data are not representative samples because students choose to take those tests. From its beginnings in the 1960s, NAEP has presented a psychometrically sound picture of student achievement in the elementary and secondary schools.

Why wouldn't we want to continue such a valuable source, especially when it informs policy-making?

Lessons in assessment issues

In November 2016, the Governing Board adopted a Strategic Vision to facilitate the greater awareness and better use of NAEP.[xiv] The LTT presents a perfect opportunity to fulfill that Strategic Vision by explaining trends in the achievement gap among various populations, and demonstrating the effects on test scores of changing demographics.

Too often, it seems a simple story based on the overall results is told in the news media and not a more complex tale considering major changes in the demography of the student test-takers. This complexity can especially be seen in the uniqueness of the LTT's administration over 45 years.

For instance, the news media will report that no significant change occurs in either reading or mathematics for all tested 17-year-olds from the early 1970s to the current decade. But, quite a different story is shown by looking at the scores of the three major subgroups composing almost all of the tested students. Over four decades, the achievement gap decreased: black and Hispanic students made academic progress while the scores of white students also increased. Everyone was a winner.[xv]

In reading the more recent releases of NAEP results, I notice that there is more emphasis on these disaggregated results and the reasons for variances from the overall results. I would like to thank the Governing Board and the National Center for Education Statistics[xvi] for those efforts to explain more fully what the test scores show.

I would go further and urge that when NAEP results are released, the overall results and the disaggregated results should be presented together on the same page. The full story is missed too often by the news media if one concentrates on the overall trends.

Since NAEP began 45 years ago, the demographic changes have been so dramatic that it is necessary to explain them at each opportunity. For example, 13-year-old white students were 80 percent of NAEP-tested students in 1978, but declined to 56 percent in 2012. Black students increased from 13 to 15 percent, while Hispanic students grew dramatically from 6 to 21 percent.[xvii]

White students as a group generally score highest of the three groups, but their percentage of all students has dramatically declined. Black students are performing better than in years past but not as high as white students, while their percentage of all students has grown. Hispanic students

are also scoring higher but again not as high as white students, while their percentage of all students has dramatically increased.

So, the scores of black students and Hispanic students went up as did their proportion of the students tested; but, the increased scores were not enough to make up for fewer white students who scored higher. The result is no general gain in test scores while below the surface there is a gain for each major subgroup.

This so-called "Simpson's Paradox" is difficult to explain to the news media and thereby to the public. So, every opportunity should be seized to do so. This phenomenon appears in both the LTT and the main NAEP, but the long-term administration of the former makes the changes more dramatic.

Why two assessments and not one?

Even given all those reasons for continuing the LTT, a nagging question is: why should there be two assessments—the LTT and the Main NAEP? Could one assessment do it all?

As Dr. Haertel points out, LTT's content is simpler and more traditional than that espoused in current curriculum reforms. The LTT assessments address "a fairly low-level traditional subset of contemporary curricular objectives." In contrast, the main NAEP "evolves in response to changing curricular priorities and expectations for schooling outcomes…"

The LTT can serve as an "anchor," as described by Dr. Haertel, since "the LTT has measured the same content for decades." Even if the content leans towards basic skills, whether students have mastered those skills and knowledge should be known.

The second point is that LTT tests by age level, and main NAEP by grade level. As Dr. Haertel points out, contrasts between age-based and grade-based gaps and trends can be useful because children are starting school at a later chronological age and students of various racial and ethnic groups are retained in grades at different rates.

Lastly, cost is a concern in retaining two assessments. Dr. Haertel has presented an option reducing costs while retaining the integrity of both sets of tests. I do not know enough about the technicalities of his proposal to endorse it fully; but if it retains the LTT as a valid, dependable assessment, I would hope that the Governing Board would seriously consider his ideas.

Conclusion

In sum, the National Assessment Governing Board should maintain the integrity and usefulness of the LTT, in spirit as well as in form. To accomplish this, the Long-Term Trend should be administered every few years.

Other advantages that would flow from that policy are that the LTT would continue to be influential in policy-making, and the Governing Board's new Strategic Vision could be further implemented.

Lastly, creativity in assessment ought to be encouraged. If Dr. Haertel's proposal for the LTT can maintain the integrity and usefulness of that assessment, then the Governing Board should consider its adoption.

From my experiences, two aspects of the LTT are the essence of what should be retained. First, the length of time is unique. LTT starts in the early 1970s while the main NAEP originates in the early 1990s; and those two decades should not be lost. The trend lines should be maintained as far back as possible. Second, what students have been tested on in the LTT over these 45 years should be retained as much as possible. As I understand it, that is necessary to maintain the integrity of the trend lines in that assessment.

Let me end as I started. I am not a psychometrician, I am a mere lawyer who has been involved in policy. If creative assessment experts find some way to retain the essence of the LTT in a simpler way than we have now, that would be to the good. But, we must retain that essence.

May I again commend the Governing Board for sponsoring this review of its decisions about the LTT. This is a very thoughtful way to proceed on making wise policies.

---

[i] The Main NAEP has also measured what students have actually attained since the early 1990s.

[ii] P.L. 107-279. Title III, the *National Assessment of Educational Progress Authorization Act.*

[iii] P.L. 100-297, the *Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988,* Title III.

[iv] Robert Rothman, "NAEP Board Fires Researchers Critical of Standards Process," *Education Week,* September 4, 1991.

[v] Mary Lyn Bourque, "A History of NAEP Achievement Levels: Issues, Implementation, and Impact. 1989-2009," Paper Commissioned for the 20th Anniversary of the National Assessment Governing Board, March, 2009.

[vi] P.L. 103-282, the *Improving America's Schools Act*.

[vii] J.W. Pellegrino et al, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational progress*, National Academy Press, 1998.

[viii] P.L. 107-110, the *No Child Left Behind Act of 2001*.

[ix] Sam Dillon, Students Ace State Tests, but Earn D's from U.S., *The New York Times,* November 26, 2005.

[x] Jim Hull, *The Proficiency Debate: At a Glance*, Center for Public Education, National School Boards Association, 2007.

[xi] Tom Loveless, *The NAEP Proficiency Myth*, Brown Center Chalkboard, Brookings Institution, June 13, 2016.

Leina Heltin, What Does "Proficient" on the NAEP Test Really Mean? *Education Week*, June 15, 2016.

[xii] Christopher Edley Jr., Judith A. Koening, *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*, National Academy of Sciences, November 2016.

[xiii] __ *No Time to Lose*, National Conference of State Legislators, August 2016.

[xiv] __ *Strategic Vision*, National Assessment Governing Board, November 18, 2016.

[xv] National Assessment of Educational Progress, 2012 Long-term: Summary of Major Findings, Institute of Education Sciences, National Center for Education Statistics, U.S. Department of Education. Data downloaded August 14, 2014 from http:/www.nationsreportcard.gov/ltt_2012/summary.aspx.

[xvi] The National Center for Education Statistics of the U.S. Department of Education is responsible for writing and releasing the reports on the National Assessment of Educational Progress.

[xvii] National Assessment of Educational Progress, 2012 Long-term: Summary of Major Findings, Institute of Education Sciences, National Center for Education Statistics, U.S. Department of Education. Data downloaded August 14, 2014, from http:/www.nationsreportcard.gov/ltt_2012/summary.aspx.