Guiding Principles and Suggested Studies for Determining When the Introduction of a New Assessment Framework Necessitates a Break in Trend in NAEP

Jeffrey Nellhaus
Massachusetts State Department of Education

Peter Behuniak
University of Connecticut

Frances B. Stancavage
American Institutes for Research

September 2009
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1070 Arastradero Road, Suite 200
Palo Alto, CA 94304-1334
Phone: 650/ 843-8192
Fax: 650/ 858-0958

Acknowledgments

The authors are indebted to Charles Blankenship for his help in compiling the data for Appendix A. In addition they wish to thank the NVS panel, and particularly Larry Hedges, for their assistance in formulating the issues discussed in this paper as well as for critical review of the manuscript in progress.

Contents

Introduction

Most educational researchers have heard the adage, "If you want to measure change in performance, don't change the measure." At the same time, however, what students need to know and be able to do may change over time as research on teaching and learning provides new insights into the educational process, science and technology advance, and expectations for student achievement rise in response to global competition. Assessments must reflect these various changes or risk becoming irrelevant. But at what point do changes in an assessment mean that it is measuring something fundamentally different than the old version, so that the results generated by it can no longer be validly compared to the results generated by previous versions of the assessment? The focus of this paper is to address this question as it pertains to reporting the results of NAEP assessments that are changed to reflect new or revised assessment frameworks.

To help answer the question, it is useful to begin with the relevant entry in the *Standards for educational and psychological testing*, in which Standard 4.16 states:

> If test specifications are changed from one version of a test to a subsequent version, such changes should be identified in the test manual, and an indication should be given that the converted scores for the two versions may not be strictly equivalent. When substantial changes in test specifications occur, either scores should be reported on a new scale or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test. (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999.)

For NAEP, the test specifications referred to in Standard 4.16 are embodied in the NAEP assessment frameworks.

Each area assessed by NAEP—mathematics, reading, science, writing, art, foreign language, U.S. history, economics, geography, and civics—has its own framework. The frameworks ensure constancy in the construct from one test administration to the next by guiding item development, test construction, and test administration and reporting policies. But as noted above, the frameworks are updated periodically to ensure that they remain consistent with advances in content knowledge and expectations for student performance.

Several NAEP assessments will be based on new frameworks in the near future: reading and science in 2009, writing in 2011, and U.S. history and civics in 2014. In response, NCES will need to decide whether the constructs embodied in the new frameworks and manifest in their corresponding assessments have changed so substantially that, in order to conform to Standard 4.16, the results generated by the new assessments will have to be reported on a new scale. The determination that the new assessment should be reported on a new scale would likely necessitate a break in the trend line.

Previous revisions to NAEP frameworks provide some context for understanding past decisions by NCES about whether or not to maintain a trend line. For instance, the NAEP mathematics framework has been revised twice since 1990. The first update, which took effect with the 1996 assessment, was made to bring the framework better in line with standards adopted by the National Council of Teachers of Mathematics. After studies were conducted to determine whether the results of the 1996 assessment could be reported on the same scale as the results of the 1990 and 1992 assessments, NCES decided to maintain the trend line at all three grades (4, 8, and 12).

The second revision of the NAEP mathematics framework took effect with the 2005 assessment. Changes made to the framework included modifications in the number of items by content area and revisions intended to improve the clarity and specificity of the content objectives at each grade level. In contrast to the decision in 1996 to maintain the trend at all three grades, the trend in 2005 was maintained at grades 4 and 8, based on the judgment that the revisions were relatively minor, but broken at grade 12, where the updated framework called for substantial changes in both the distribution of items by content area and the policy on use of calculators.

Appendix A and figure A-1 summarize changes in NAEP frameworks that have occurred since 1990 and briefly describe decisions about trend lines that accompanied the changes.

Purpose

The purpose of this paper is to recommend guiding principles, suggested studies, and decision-making processes to NCES leadership to assist them in determining whether the results generated by an assessment based on a new NAEP framework can be validly reported on the same trend line as previous versions of the assessment. It has become increasingly important that NCES adopt sound policies and procedures for reporting results of assessments based on new or updated assessment frameworks, as NCLB has created heightened interest in NAEP among many audiences. Moreover, many state assessment programs are updating their content standards and test specifications and may model their decisions relating to trend lines after NAEP.

We begin with a discussion of principles that NCES could follow in making decisions about whether or not to break an existing trend line. This is followed by suggestions for the types of studies and decision-making procedures that might be found useful for such deliberations.

Guiding Principles

- **A substantial change in the construct necessitates a break in the trend**. The construct measured by each NAEP assessment is embodied in the corresponding NAEP assessment framework. The frameworks describe what should be tested, how it should be tested, and what constitutes basic, proficient, and advanced performance. Accordingly, when a new framework is introduced or an existing framework is modified, and new test items are developed and

administered to reflect the new or revised framework, constancy in the trend can no longer be assumed. Rather, evidence needs to be gathered to determine whether the construct has changed so substantially that the results of the assessment need to be reported on a new scale, with new achievement levels, and the trend line broken.

- **Clear criteria should be established for determining whether the construct has changed substantially or not.** The criteria should involve evidence of at least two types: (1) whether the content frameworks of the two assessments are essentially the same or not, and (2) whether the items based on the new framework are measuring the same construct as were the items based on the old framework. For each of these types of evidence, specific criteria on the goodness of fit between the old and new assessments should be decided *before* gathering the evidence. The kinds of studies that might be done are described in Section III.

- **Evidence gathering and decisions made on the basis of the evidence should be public, systematic, well documented, and inclusive of input from key stakeholders**. While NCES and its contractors are responsible for designing and conducting the relevant studies and analyses, there should be a role for external stakeholders to review study designs. Potential roles for stakeholders should also include participating in studies (especially studies where content expertise is necessary, such as comparisons of frameworks and test items), as well as assisting in determining the extent to which the body of evidence gathered from the various studies and analyses indicates a substantial or relatively minor change in the construct being measured.

- **Bridge studies should be undertaken when a decision is made to break the trend.** If there is a decision to break the trend, it is important that bridge studies be undertaken in order to know how students taking the new assessment would have done on the old assessment.

Studies and Analyses

A. Studies involving comparisons of old and new assessment frameworks. NAEP assessment frameworks share a number of common components, including (1) subdomains, including content areas and cognitive categories, that often form the basis for reporting subscale results, (2) test specifications (e.g., number of items by content area), (3) test administration policies, and (4) achievement level descriptions. Accordingly, studies of the frameworks should include:

- **A comparison of the content and cognitive categories**. Each NAEP assessment framework defines the domain (content) to be assessed. The domain is typically broken up into subdomains, which may be cross-cutting. For example, the mathematics framework contains "content objectives" grouped into "content areas," and it also calls for the assessment of certain "cognitive processes" (National Assessment Governing Board, 2008a) The subdomains described by the reading framework include "vocabulary,"

"passage type," and "cognitive targets" (National Assessment Governing Board, 2008b). The science framework describes "topics" grouped into "major fields," and it further specifies that certain processes of "knowing and doing science" be assessed (National Assessment Governing Board, 2008c).

Therefore, studies that compare the content of any two frameworks should be designed to capture the extent to which the subdomains within the two frameworks are the same or different. To the extent possible, such studies should provide quantitative information. For example, in mathematics, the study should indicate the percentage of content objectives that are the same or different. In reading, it would be desirable to know what percentage of passages have the same classification, and in science, the percentage of topics within fields that have the same classification.

- **A comparison of the test specifications.** To understand the extent to which the old and new frameworks specify assessments that emphasize similar areas of the construct, this study should compare the number/proportion of test items that each framework specifies by content area. Any changes in specifications relating to item types (e.g., multiple-choice, constructed response, writing prompts) should also be examined and quantified. In the case of reading tests, the comparison should include changes in specifications regarding type of text (e.g., literary, informational) and genre (e.g., prose, poems, drama).

- **A comparison of test administration policies.** This study should compare policies relating to the use of manipulatives, the use of technology (e.g., calculators for computation and computers for word processing), test participation requirements, and test accommodations for students with disabilities and English language learners.

- **A comparison of achievement level descriptions.** This study should compare the language in each framework (old and new) used to describe student performance at the various NAEP achievement levels (Basic, Proficient, and Advanced). When changes in achievement level descriptions clearly indicate higher (or lower) expectations for attaining any given performance level, serious consideration should be given to resetting performance level cut scores. If it is determined that policy interests will be best served by maintaining trends in performance level results—either by using existing cut scores, or by identifying new cut scores through methods such as equipercentile equating—modifications should be made to the new achievement level descriptions, as required, to ensure that they are consistent with the performance students demonstrate at each level.

For each of these studies, NCES should establish statistical criteria that indicate the goodness of fit between the old and the new assessment component (e.g., content objectives, administration policies) before the studies are undertaken.

B. Studies involving comparisons of old and new item pools

- The extent to which the constructs defined by the old and new frameworks are substantially the same or different can also be studied by determining the extent to which old items and new items (and old and new passages in the case of reading tests) "fit" their "parent" frameworks as well as each other's frameworks.

- Comparisons of the item pools are arguably more important than comparisons of the various components of the framework because, even when frameworks may contain differences, the items generated from them may be judged to address similar knowledge, concepts, and skills.

- Similar to studies designed to compare old and new frameworks, studies designed to compare old and new item pools should provide information broken down by subdomain, including both content area subdomains and, when applicable, cognitive processes. The primary measure for characterizing the extent to which the item pools are the same/different should be the percentage of items that "fit" each framework, broken down by subdomain.

- To reduce bias, studies of this nature should be carried out with a combined pool of items where the identity of the items as being old or new cannot be determined.

C. Studies based on empirical data derived from live tests and field tests

- Empirical data derived from administration(s) of the old and new items should be used to compare student performance over the entire domain, for the various content areas defined by the framework (e.g., literary text vs. information text), and for various student subpopulations. Population invariance should be a key factor in determining constancy in the construct (Dorans & Holland, 2000).

- One model for conducting these empirical studies is to carry out a braided bridge in which equivalent groups of students are given either: 1) two old blocks, 2) two new blocks, or 3) one old block and one new block. For the sake of efficiency, it may be possible to use data from a recent live administration of the old assessment to examine the two-old-block condition.

- The braided bridge study should address the following questions (Moran and Xu, 2009):
    - What differences are observed in the difficulty, response rates, speededness, and reliability of old and new item blocks?
    - How similar are the IRT calibration results when new and old items are scaled jointly or separately?
    - What are the relationships between the old and new subscales (as evidenced by correlations between scale scores)?

  - How similar are the group-level scale score results for various subgroups, based on the old and new assessment instruments?

- Other analyses, such as factor analysis suitable to the NAEP methodology, should be considered as a means for evaluating the stability of the measured constructs.

- Statistical criteria should be determined a priori to guide the decision of whether the magnitude of differences exposed by these studies indicates a relatively minor or more substantial change in the construct.

Process for Making a Preliminary Determination

NCES should convene a group of stakeholders and involve them in a systematic review of the findings generated by the various comparative and empirical studies. The review should result in a recommendation or report by the stakeholder group that can be used by NCES, along with other information that NCES may want to consider, to make a *preliminary* decision about whether to maintain or break a trend line for a NAEP assessment based on a new or revised assessment framework.

Validation Studies

The preliminary decision about whether to maintain or break the trend line should be validated using data derived from the first administration of the new assessment, including data from any bridge studies conducted as part of that administration. Validation studies should include many of the same analyses suggested earlier in this paper.

Schedule

Assuming that a preliminary decision about whether to maintain or break the trend line will need to be made prior to the first administration of the new assessment (with a final decision to be made as soon as possible thereafter), the studies suggested in this paper should be carried out in a timely fashion as follows:

- Comparative studies of the various components of the current and new assessment frameworks should be conducted immediately after the new framework is adopted.

- Comparative studies of test items should be conducted after the new item pool has been field-tested and the items suitable for future live administrations of NAEP have been identified.

- Empirical studies using field test data and including old and new items should be conducted as soon as these data are available.

- Studies designed to validate preliminary determinations should be conducted immediately after the first administration of the new assessment.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Dorans, N. J. & Holland, P. W. (2000). Population invariance and equitability of tests: Basic theory and linear case. *Journal of Education Measurement*, *37*, 281-306.

Moran, R. &, Xu, X. (2009). Maintaining Trend Measurement across Assessment Frameworks: Psychometric Design Issues. Paper presented at the annual meeting of the AERA, April 13-17, San Diego, CA. Unpublished work by ETS, Princeton, NJ.

National Assessment Governing Board (2008a). Mathematics Framework for the 2009 National Assessment of Educational Progress. Washington, DC.

National Assessment Governing Board (2008b). Reading Framework for the 2009 National Assessment of Educational Progress. Washington, DC.

National Assessment Governing Board (2008c). Science Framework for the 2009 National Assessment of Educational Progress. Washington, DC.

Appendix A – Summary of Changes to NAEP Assessment Frameworks and Trends

The history of explicit frameworks, as currently employed by NAEP, started with the 1990 mathematics framework. At that time the National Assessment Governing Board (NAGB) had taken over responsibility for determining the content to be tested, and the frameworks were developed by NAGB to guide the test developers and to explain the assessment to stakeholders. It was expected that frameworks would remain current for approximately a decade, after which new frameworks would be written and new trend lines established.

Up to that time, NCES and ETS had used bridge studies to create overlapping trend points when major changes were made in the assessment.[1] Generally speaking, however, the necessity of breaking the trend line was presumed rather than evaluated empirically. Other approaches to dealing with changes in the assessments included reanalyzing old assessments to fit new trend lines retrospectively. This occurred, for example, with the 1992 mathematics assessment, when PARSCALE was introduced and partial credit scoring became possible. The 1990 mathematics assessment was reanalyzed by the new methods, and a trend line was established from 1990 onward.

Figure A-1 shows the history of NAEP assessments since 1990, with framework changes and trend decisions noted. New frameworks continued to be written in different subject areas throughout the 1990s, and NAGB has commissioned periodic reviews of the established frameworks in various subject areas. The review committees have been charged with recommending whether major or minor changes are needed to keep the frameworks current.

Prior to the 2009 assessment, there have only been five instances in which NAGB-era frameworks were modified. One of these modifications (the writing framework for the 1998 assessment) was driven by psychometric concerns because it had not proved possible to scale the 1994 writing assessment using subscales, as specified by the 1994 framework. The other modifications were in the mathematics frameworks used for the 1996 and 2005 assessments, the reading framework used for the 2003 assessment, and the U.S. history framework used for the 2006 assessment.

In all but one of these cases (grade 12 mathematics in 2005), NAGB determined, based on the recommendations of their expert panels, that the content changes were not sufficient to require a break in trend. NCES and ETS undertook confirmatory analyses to test the validity of the trend continuation in 1996: the 1996 mathematics assessment was scaled with and without the new items, and the assessment results were not reported until these analyses had been completed and judged to uphold the continuation of the trend line. This was the only pre-2009 trend decision that was evaluated psychometrically, however, beyond the normal QC analyses carried out as part of the scaling process.

[1] In fact, long term trend began as a bridge study when NAEP assessments changed in the mid-1980s, but it then became institutionalized as an on-going and parallel trend line.

Three framework revisions were put into place for the 2009 assessment cycle—in reading, in mathematics at grade 12, and in science. Initially, each of these was judged to be of sufficient magnitude to require a break in the trend line. However, with agreement from NAGB, NCES subsequently undertook a series of empirical studies designed to evaluate whether or not it was appropriate to continue the existing trend lines in reading and mathematics. These studies included content comparisons of the frameworks and item pools as well as empirical studies using a braided bridge. The braided bridge design also allowed for a one-year overlap between the old and new trend lines if continuation was not appropriate. For science, on the other hand, the new framework was implemented without a bridge study and without any analyses to evaluate whether or not it would be more appropriate to continue the trend line.

Figure A-1. NAEP Assessment Frameworks and Trends Since 1990

	1990	1992	1994	1996	1998	2003	2005	2006	2009
Reading		1st Framework adopted				Modified to update concepts of reading. Trend maintained.			Replaced. Studies ongoing to determine trend.
Math	1st Framework adopted			Modified to align with NCTM. Trend maintained.			Modified to update content. Trend broken in grade 12.		Replaced at grade 12. Studies ongoing to determine trend.
Science				1st Framework adopted					Replaced. Trend broken.
Writing		1st Framework adopted			Modified. Trend broken.				
US History			1st Framework adopted					Modified to update content. Trend maintained.	
Geography			1st Framework adopted						
Civics					1st Framework adopted				