Future of NAEP Long-Term Trend Assessments[1]

A white paper prepared for the National Assessment Governing Board

Edward Haertel, Ph.D.
Stanford University

December 9, 2016

The National Assessment of Educational Progress (NAEP) encompasses several distinct assessments. In addition to its main series of periodic assessments at grades 4, 8, and 12 in various subject areas (main NAEP), these include the NAEP Long-Term Trend (LTT). The National Assessment Governing Board (Governing Board), created by Congress to formulate policy for NAEP, strives to minimize administrative and testing burdens entailed by NAEP data collections and to realize efficiencies in NAEP operations. Accordingly, the Governing Board's Strategic Vision calls for an examination of policy and technical implications related to the future of NAEP LTT assessments in reading and mathematics. A stakeholder outreach event, to be held early next year, will inform the Governing Board's deliberations as to whether LTT assessments should be continued independently from main NAEP assessments, whether it is feasible to blend LTT assessments with main NAEP assessments, and related questions. This white paper has been commissioned by the Governing Board in preparation for that event. It is intended as both a summary of the history of the LTT and an analysis of some options that the Governing Board and the National Center for Education Statistics (NCES) may wish to consider.

Introduction

The National Assessment of Educational Progress (NAEP) was conceived in the early 1960s as a new kind of assessment program, designed to describe and track the academic skills of United States citizens of school age and young adults. It was bold, innovative, and experimental. NAEP assessments were conducted beginning in 1969, but by the end of the 1970s, a decade later, some compromises had been made. The initial plan for NAEP had been scaled back considerably, some aspects of the original NAEP design had been reworked, and substantial further changes appeared unavoidable. Thus, around 1982, a Request for Proposals was issued for a new NAEP contractor, and under the new contractor, NAEP was redesigned and NAEP reporting scales were introduced for the first time. The first data collections under the new design occurred in 1984. The changes made in 1984 were so significant that it was not clear how, or even if, meaningful comparisons could be made between students' performance on the redesigned NAEP and the performance of earlier student cohorts on assessments dating back to NAEP's beginnings. In order to maintain comparability, NAEP continued with two parallel data collections. A scaled-back version of the original NAEP data collection design was continued, under the name "NAEP Long-Term Trend" (LTT). The new assessment design, data collection, analysis, and reporting came to be called the "main NAEP."

At the same time as the LTT tracked students' performance based on reading and mathematics learning objectives dating back to the 1970s, various aspects of main-NAEP assessments, including the subject area frameworks guiding main NAEP content, rapidly evolved. Around 1990, significant changes to the main-NAEP assessment frameworks for both reading and mathematics required that new score scales be

[1] This paper has benefited enormously from constructive comments and corrections by numerous reviewers. Their assistance is gratefully acknowledged. Any remaining errors are the responsibility of the author. Views and opinions expressed are the author's own and do not necessarily represent those of the U.S. Department of Education, the National Center for Education Statistics, or the National Assessment Governing Board.

created for main-NAEP reporting, so that today, the term "main NAEP" is used to refer to these NAEP assessments from 1990 to the present. At the same time, the LTT has preserved a capability for making direct comparisons of reading and mathematics assessment results from the earliest NAEP data collections (in 1970-71 for reading and 1972-73 for mathematics) up to and including the most recent LTT data collections in 2012.[2]

Thus, although main NAEP also provides information about achievement trends[3] over time, the main-NAEP trends extend back only as far as 1990 for mathematics[4] and 1992 for reading, and they are reported on different measurement scales from those still used for the LTT. Where they cover the same time periods, the LTT trend lines and the (shorter) trend lines from main NAEP do not quite match up (Beaton and Chromy, 2010).

At this point, the future of the LTT is unclear. LTT data collections planned for 2016 have been postponed twice, first to 2020 and then to 2024, primarily for budgetary reasons as NAEP has responded to other priorities.

The focus of this white paper is on the LTT assessments in reading and mathematics, as well as their relation to the main-NAEP assessments in these same subject areas. It is intended as a starting point for a broad discussion of the LTT assessments, offering an overview of issues and options that might be explored in greater depth in future papers and symposia. Specifically, this white paper offers a brief history of the LTT and then addresses the following questions:

- What are some arguments for and against continuing the LTT component of NAEP in essentially its current form versus dropping it altogether?
- How might the LTT component instead be integrated (or blended) with main-NAEP assessments?
- How might historical LTT data, main-NAEP data, and bridge study data be integrated to make NAEP more useful for longitudinal research?

History and Context of the Long-Term Trend Assessment

Planning for NAEP began when foundation support was secured in 1963 for an Exploratory Committee for the Assessment of Progress in Education (ECAPE), re-formed as the Committee for the Assessment of Progress in Education (CAPE) in 1968. In 1969, CAPE enacted a Memorandum of Understanding

[2] LTT trend lines were disrupted in 2004 when LTT data collection and analysis procedures were updated. Various changes to the LTT were made at that time, including new testing accommodation policies that brought the LTT into conformity with main-NAEP practices and with the requirements of the Individuals with Disabilities Education Act of 1990 and other legislation (see Olson and Goldstein, 1997). LTT trend lines were preserved by conducting two parallel sets of LTT data collections in 2003-04, one following prior procedures and the other following new procedures. Results from both versions, which differed by just a few points, were reported on the same scales, already established for the LTT. Those scales have been maintained for LTT assessments since 2004.

[3] Although the terminology of "trends over time" is commonly used and almost unavoidable, it must be remembered that each NAEP assessment is cross-sectional — no individual persons or schools are tracked over time. Thus, "trends" refer to comparisons of distinct groups, constituted from cohorts reaching specified age or grade levels in different years, such as 9-year-olds in 1971 versus 9-year-olds in 1975. In discussing assessment results, this report avoids referring to "changes over time" for the same reason, referring instead to comparisons of different cohorts over time.

[4] The grade 12 trend line in mathematics was disrupted in 2005, when two sets of results were reported.

(Resolution of Transfer) whereby the Education Commission of the States (ECS) assumed responsibility for NAEP governance. The ECS work was supported by an initial $1 million grant from the U.S. Office of Education (USOE), with further USOE support to follow. Thus, NAEP was born (Jones, 2004).

The assessment program envisioned by NAEP's founders looked very different from NAEP today. Originally, NAEP aspired to assess samples from the full populations of U.S. residents at ages 9, 13, and 17, as well as young adults. Ten subject areas were to be assessed.[5] As Jones (1996, p. 15) recounts, "The goals of NAEP were to report what the nation's citizens know and can do and then to monitor changes over time using objective-referenced assessment, a close kin to criterion-referenced assessment as had been proposed by Glaser and Klaus (1962) and by Glaser (1963)." Importantly, the original designers intended that "NAEP would assess knowledge and skills that could be gained from any source, not just from school learning. What citizens know can be measured; … how they acquired their knowledge would be far more resistant to discovery" (Jones, 1996, p. 15).

Lists of objectives in each subject area would be developed by a consensus process with broad representation of diverse points of view, and NAEP exercises[6] would be keyed to these objectives. There would be no such thing as a NAEP test score, or even a NAEP score scale. Instead, reporting would be in terms of estimated proportions of populations and subpopulations able to answer each exercise correctly. Some exercises would be released after each assessment, so that anyone could see examples of just what it was that respondents had been asked to do. Short-answer questions would be favored over multiple-choice. Some exercises would be given to small groups of examinees to solve together. NAEP samples would include children not in school, and these children would be tested individually, typically in out-of-school settings. No results would be reported for individuals, schools, school districts, or states, only for the nation and broad geographic regions (Northeast, Southeast, Midwest, and West); for broad demographic groups (defined by gender, ethnicity, level of parental education, degree of urbanization [size and type of community]); and for cross-classifications of such categories (Beaton and Johnson, 2004). This was a bold, radically new way of thinking about large-scale assessment, an experiment that had never been tried before on such a scale (Jones, 1996). The first NAEP assessments, in the areas of science, citizenship, and writing, took place in 1969-70. Reading was first assessed in 1970-71 and mathematics was first assessed in 1972-73 (Jones and Olkin, 2004, p. 562).[7]

The NAEP funding mechanism was changed from a grant to a contract in 1973. ECS continued to manage NAEP with USOE support until 1983, when ECS' final five-year continuation contract expired. Educational Testing Service (ETS) won the next NAEP contract, promising "A New Design for a New Era" (Messick, Beaton, and Lord, 1983), and assumed responsibility for NAEP in March 1983 (Jones, 2004). As discussed below, the ETS redesign did indeed bring major changes in NAEP's content specifications and exercises, as well as new procedures for NAEP sampling, administration, data analysis, and reporting. It would be a mistake, however, to imagine that NAEP prior to the transition from ECS to ETS was not already changing. Almost from NAEP's very beginning, it had proven necessary to modify one after another of its original guiding principles. Among other compromises, inclusion of children with severe disabilities proved prohibitively expensive. Contractors submitted fewer very easy items and a

[5] The original 10 subject areas were reading, writing, mathematics, science, literature, social studies, citizenship, art, music, and career and occupational development (Jones, 1996, p. 15).
[6] Items on NAEP assessments are sometimes referred to as "exercises," especially in older publications.
[7] For further detail, see also https://nces.ed.gov/nationsreportcard/about/assessmentsched.asp for a list of all NAEP assessments from 1969-70 to the present.

greater proportion of multiple-choice questions than had been specified. Within a few years, budgetary constraints forced discontinuation of testing for out-of-school 17-year-olds and of young adults (Jones, 1996), although 16- to 25-year-olds were assessed under a separate grant in a 1985 "Young Adult Literacy Study." Testing in areas relying most heavily on performance assessment, including art and music, also was judged too expensive to maintain. The content areas of reading and literature were combined into one in 1979-80.

Evolution of NAEP Objectives and Exercises Prior to 1983

In considering the future of the LTT, the question of what the early NAEP assessments actually measured is of some importance. NAEP exercises were keyed to NAEP objectives. Thus, it may be helpful to consider the early NAEP objectives in some detail.

ECS did not have the in-house capacity to develop objectives for all the content areas NAEP was to assess, and so it subcontracted with different organizations for objectives development in different areas (Lehmann, 2004). Moreover, to obtain different points of view, two or three separate contractors were engaged to work independently on objectives development for some areas. Thus, language arts objectives were drafted by ETS, reading objectives by Science Research Associates (SRA), and mathematics objectives by ETS, by SRA, and by the Psychological Corporation (PC). After objectives were submitted by subcontractors, a sometimes contentious process ensued, involving multiple reviews and revisions (Lehmann, 2004). Perhaps not surprisingly then, the organization of objectives differed from one subject area to another. Objectives were reviewed by lay panels as well as subject area specialists and measurement specialists. According to one knowledgeable pair of observers, however, the final result of this elaborate, broadly consultative process was that "National Assessment had not produced a set of new objectives … [but instead] its objectives represented a reorganization, restatement, and something of a summarization of objectives which frequently had appeared in print in the last quarter century" (Merwin and Womer, 1969, p. 316, as quoted by Lehmann, 2004, p. 65). It should be noted that although the exercises keyed to objectives were also carefully reviewed, apart from some tryouts to judge clarity of instructions and exercise difficulties, there was little or no pilot or field testing of exercises prior to live administration.

The objectives guiding the first (1972-73) NAEP mathematics assessment were published in 1970. There were 47 objectives for age 9, 45 for age 13, 78 for age 17, and 38 for young adults. For this first mathematics assessment, a three-dimensional scheme was specified for classifying objectives. The first dimension was "use of mathematics" (social mathematics, technical mathematics, academic mathematics, each with further subclassifications), but in the end, little use was made of this dimension, and it was dropped after the first assessment. The "objectives and abilities" dimension included "recall and/or recognition of definitions, facts and symbols," "perform mathematical manipulations," "understand mathematical concepts and processes," "solving mathematical problems — social, technical, and academic," "Using mathematics and mathematical reasoning to analyze problem situations, define problems, formulate hypotheses, make decisions, and verify results," and "appreciation and use of mathematics." These levels were reminiscent of the six broad levels of the Bloom Taxonomy (Bloom, Englehart, Furst, Hill, and Krathwohl, 1956), which were: knowledge, comprehension, application, analysis, synthesis, and evaluation. The "content" dimension listed 17 areas, including "attitude and interest." Today, almost a half-century later, some of the specific listed objectives seem quite dated, and others seem out of place in a mathematics assessment. Objectives for young adults included "use of computers," "keeping a checkbook," and "detecting flaws in advertising or propaganda arguments." One objective for 17-year-olds was "knowledge of scientific units, such as: calorie, B.T.U., foot-pound, ohm,

ampere, volt, watt, coulomb, erg, dyne, poundal, lumen, foot-candles, roentgen, angstrom, light-year, nail sizes, wire gauge, horsepower." The 17-year-old objectives also included "slide rule computation" and "explaining the long division algorithm in terms of successive subtractions," as well as "using nomographs," "finding square roots," and "computation with logarithms."

The first NAEP reading assessment was conducted in 1970-71, based on a list of objectives also published in 1970. Although the layout and organization of objectives for reading were quite different from those for mathematics, the reading objectives were also organized into broad categories reminiscent of the Bloom Taxonomy: "comprehend what is read," "analyze what is read," "use what is read," "reason logically from what is read," "make judgments concerning what is read," and "have attitudes about and an interest in reading." These categories were divided and subdivided into long lists of specifics. Within "comprehend what is read," for example, objectives included "read individual words," "read phrases, clauses, and sentences," and "read paragraphs, passages, and longer works." These in turn were divided into many subobjectives, each accompanied by lists of illustrative behaviors, down to such particulars as "interpret the sound patterns (intonations) suggested by punctuation marks" and "interpret intonation patterns such as tone of voice which are not completely represented in writing." Figures of speech to be comprehended included simile, metaphor, personification, hyperbole, litotes, metonymy, and synecdoche. The objective "obtain information efficiently" included as subobjectives "skim a paragraph or passage," "use the various parts of a book as aids in finding what is needed" (with parts such as title page, preface, index, and glossary listed by way of illustration), "find information efficiently in a variety of reference tools," and "obtain information from 'nontextual' sources." Illustrative behaviors included such specifics as use of the *Reader's Guide*, *International Index*, etc., as well as "libraries and card catalogs." As with the mathematics objectives, "recognize propaganda" was included, and another objective called for readers to "recognize the rhetorical techniques of the demagogue."

Objectives changed with each successive NAEP administration.[8] After the first mathematics assessment, roughly half the exercises were released. The remaining half were reused in 1977-78 to enable comparisons between student cohorts tested in 1972-73 versus 1977-78. Additional exercises for the second mathematics assessment were based on a streamlined set of objectives, dropping the "use of mathematics" dimension and using fewer, broader categories for the content and process dimensions. For "content," just five categories were used, down from 17 in the first assessment. These were "numbers and numeration," "variables and relationships," "shape, size, and position," "measurement," and "other topics." Just four "process" categories were used, down from six. These were "mathematical knowledge," "mathematical skill," "mathematical understanding," and "mathematical application." These changes were made in part with an eye toward grouping exercises in a way that would be meaningful for reporting NAEP findings. The second mathematics assessment also introduced student use of calculators on some exercises. Approximately one-third of the exercises from the second NAEP mathematics assessment were released, with the remaining (nonreleased) exercises again carried forward for reuse to measure differences in performance of student cohorts over time. For the third NAEP mathematics assessment, both content and process categories were again reworked, although changes were not as extensive as those from the first to the second mathematics assessments. Among other changes for the third mathematics assessment, "technology" was added as a content category and objectives addressing "estimation" ability also appeared.

[8] For mathematics, further information concerning changes to exercises and objectives from the first (1972-73) to the second (1977-78) to the third (1981-82) assessments is provided in NAEP Report No. 13-MA-10, *Mathematics Objectives: 1981-82 Assessment* (NAEP, 1981a).

After the first NAEP assessments of reading and literature as two separate areas in 1970-71, reading was assessed again in 1974-75, and reading and literature were combined and assessed as a single area in 1979-80, under a data collection design that also enabled separate reporting of achievement in reading. As with mathematics, roughly half of the reading exercises from the 1970-71 assessment were released, with the remainder kept nondisclosed for future use. The *Procedural Handbook* for the 1979-80 reading and literature assessment summarized the evolution of the reading objectives over prior assessments as follows:

> The first reading objectives were … [c]omprehensive in scope and very detailed, [and] addressed literary as well as nonliterary texts, [and] literary terms and skills as well as terms and skills more closely associated with reading instruction … In contrast, the second reading objectives … were much less detailed and concentrated solely upon the goals of reading instruction defined quite narrowly. Consultants felt that reading should be differentiated from literature since each was a separate assessment area and a separate instructional field in the schools. The second objectives were also somewhat more behaviorally oriented and more directly tied to what might be measurable.
>
> The first [1970-71] literature objectives … stressed knowledge of classic works, skills necessary for interpreting works and activities that promote involvement with literary experience. They ignored skills involved in learning to read. The objectives developed for the 1975-76 literature assessment (which, for financial reasons, never took place) were quite different. Rejecting the notions that "literature of excellence" could be defined or that acquaintance with classics could be assessed meaningfully, the consultants placed more emphasis on response and valuing. Instead of defining literature as "great books," they defined it as "language used imaginatively" and created objectives designed to determine how much exposure students have had to imaginative language in a number of social and academic contexts. Again, the objectives made no mention of reading skills per se. (NAEP, 1981b, pp. 1, 4)

The 1970-71 reading objectives, the 1974-75 reading objectives, the 1970-71 literature objectives, and the 1975-76 literature objectives all informed the development of reading and literature objectives for the 1979-80 reading and literature assessment. For that assessment, objectives were organized into four areas: "values reading and literature," "comprehends written works," "responds to written works," and "applies study skills in reading." After considerable discussion, it was concluded that word attack skills would be omitted. Reading rate, skimming, and scanning were assessed "in a limited way" as part of study skills (NAEP, 1981b, p. 5).

The 1981-82 NAEP mathematics objectives and the 1979-80 NAEP reading objectives have continued to define the content of LTT assessments in these respective subject areas. Thus, they remain highly relevant to the problem of describing the domains of knowledge and skills addressed by these assessments. This topic is addressed in greater detail at the end of this discussion of history and context, in a subsection titled "What do the LTT assessments measure?"

Evolution of NAEP Reporting Prior to 1983

As already noted, early NAEP measurement targets were akin to observable attributes, as opposed to underlying (latent) attributes (constructs) accounting for observed regularities in performance (cf. Kane, 2006, p. 32). This objectives-based vision comported well with the language and logic of behavioral

objectives and promised a direct, immediate form of interpretability via actual exercises together with information about respondents' success rates. In accordance with this vision, the original model for reporting NAEP results was to provide proportions correct for individual exercises together with publication (release) of a subset of those exercises. Lee J. Cronbach, one of the original ECAPE (later CAPE) members, recalled in a 2004 interview that:

> First, … there was interest in making the results comprehensible to teachers in terms that they could put to use in their teaching. Second, the reports to the public were to be as free as possible from measurement technology … The idea of scores was dismissed pretty much outright. … In the end, we thought we could establish something a bit like opinion polling … [with] newspaper releases and have columns in which they laid out the questions and the data that came back. … [T]he idea was to directly report the exercises together with the percentages correct. It might have been presented first for all students, and then for whatever subgroups were to be identified. (Cronbach, 2004, pp. 141-142)

Unfortunately, this reporting model proved awkward and ultimately unworkable for two reasons. First, reports of proportions correct for hundreds of separate exercises proved difficult to comprehend and summarize. Second, exercise-level reporting absent any kind of item calibration made it difficult to compare successive cohorts over time. To address the first of these concerns, shortly after the first NAEP assessments, exercise-level reporting was supplemented with reports of average performance for groups of exercises, often sets of exercises keyed to the same objective or a group of related objectives.[9] However, reports of average percentages correct for groups of exercises also had serious limitations. Because the exercises released after each assessment were not reused, new exercises had to be added after each assessment to replenish the pool. Adding exercises meant that the set of exercises keyed to a given objective was shifting over time, and so average performance one year, across one set of exercises keyed to a given objective, was difficult to compare with average performance for some other year, across a different set of exercises keyed to that same objective (Beaton and Johnson, 2004; Jones, 1996). Note also that, while individual exercises could each provide a series of data points over the assessments for which they were used, as soon as an exercise was released, its particular series of data points was terminated, because released items were never again used. The dwindling pool of secure exercises dating back to the earliest assessments was one of the major problems prompting calls for a new NAEP design in the early 1980s.

Cronbach went on to recall that:

> The only statistic that was of great interest to us was the standard error and the percentage correct on each item in turn. Well, the whole point, you see, was to get away from reports about "people are proficient," "people are meeting the national objectives," or whatever. … Our position from the beginning was that we were going to offer minimally processed reports on what was observed in student performance. We did not expect any user to need technical understanding to deal with the reports. Obviously, we had this terrible high-bandwidth reporting scheme that collapsed of its own weight. Nobody took it seriously. But this was still the basis for deciding that an assessment exercise was satisfactory. (Cronbach, 2004, pp. 147-148)

[9] Exercise-level proportions correct were still included in appendices to NAEP reports.

In a paper commissioned for the 20th anniversary of the Governing Board, Lawrence C. Stedman recounted the ECS plan for "minimally processed reports," which in some ways followed the earlier ECAPE/CAPE vision:

> In 1970, ECS explained that it would "issue National Assessment reports from time to time without interpreting the results or explaining their implications" ... Instead, ECS routinely asked subject matter professional associations … to independently write interpretive commentaries ... Early NAEP reports were often written by panelists (sic) of leading math, social studies, or literacy educators and included quotes and observations from them about the diverse meanings of the findings ... Many perspectives were heard. (Stedman, 2009, p. 36)

Given such an unwieldy reporting scheme for even a single assessment, it is perhaps unsurprising that apparently little or no attention was devoted to the problem of describing changes across two or more assessments. As Cronbach (2004, p. 153) lamented, "You can collect information on a tremendous number of items. But when you try to aggregate items and say these items all require [for example,] understanding of measurement principles, that is so remote from the subtle item itself that most of the information is getting lost because the different questions will have different meanings. … But you can't run an assessment at that fine grain."

1983 NAEP Redesign

In March 1983, ETS took over the management of NAEP from ECS, promising "A New Design for a New Era" (Messick, Beaton, and Lord, 1983). The major conceptual shift in the NAEP redesign was a move from objectives-referenced to construct-referenced assessment. Under the original vision for NAEP, each particular exercise had its own story to tell. Under the new design, exercises were interchangeable indicators of examinees' standing with respect to underlying traits or constructs. In retrospect, this shift was probably inevitable. It was in one way an advance because by the 1980s, the influence of behaviorist psychology was waning. The conception of educational goals in terms of behavioral objectives, epitomized by the Bloom Taxonomy published in 1956, was also giving way to conceptions framed by the cognitive psychology of school subjects, with construct-based theorizing about the structure of knowledge in long-term memory, the relation of new knowledge to what was already known, and metacognition. It was in another way a retreat, however, because the move from objectives-based to construct-based assessment made NAEP much more like a conventional test, measuring just a few broad constructs and inviting primarily norm-referenced — not criterion-referenced — interpretations, albeit at the level of populations, not individuals. The description of populations in terms of score distributions, even multivariate distributions across some small number of constructs, seemed impoverished relative to the rich portrait promised by information about performance across hundreds of specific exercises keyed to dozens of specific learning objectives.

Item Response Theory (IRT) was at the heart of a new, sophisticated, and state-of-the art data analysis plan used to create reporting scales (NAEP scale scores) for each subject area and to estimate scale score distributions for various populations and subpopulations. Creation and use of scale scores solved some serious problems that the original NAEP design had faced, including the problem of maintaining comparability over time while refreshing the exercise pool.

IRT also promised to help offset the loss of criterion-referenced information about performance on individual exercises, offering instead a new form of criterion-referencing whereby released exercises could be pegged to particular scale scores to illustrate what kinds of things students at different score

levels knew or were able to do (Beaton and Allen, 1992).[10] In addition to illustrating points along NAEP scales with individual items, more general descriptions of scale score regions were also created. For this purpose, a set of "performance levels" was chosen, equally spaced along a NAEP score scale. Next, sets of exercises were identified that, by and large, students at each specified performance level could answer correctly but those at the next lower performance level could not. Finally, content experts reviewed these sets of exercises and wrote descriptions of the knowledge and skills that students at each performance level could demonstrate.[11]

The ETS scaling model for NAEP relied on data collections following a "balanced incomplete block" design, with exercise booklets "spiraled" (the BIB-spiral design). That meant that exercises were first assembled into blocks, and two or three blocks of exercises were then assembled into booklets, in such a way that different combinations and orderings of blocks appeared together in different booklets. Each exercise block appeared in all possible positions (beginning, middle, or end) in one booklet or another, and all possible pairs of blocks appeared together in at least one booklet, but not all possible block combinations and orderings were used. Thus, this was an "incomplete" block design. The different booklets were then "spiraled," meaning that they were assembled (physically stacked) in rotation starting at different points in a sequence, so that when booklets were distributed within a single classroom, dealt off the top of the pile, nearly equal numbers of children within each testing session would receive each possible test form. This highly structured data collection design was intended to assure that every test item and every pair of test items would be administered to randomly equivalent subsamples of students — nearly the same numbers of students and the same types of students for every subsample — so that intercorrelations between all possible pairs of items could be calculated.

NAEP's transition to IRT scaling did not go smoothly (Mislevy, 1987). Beaton, writing about the analysis of 1983-84 reading data, summarized some of the difficulties:

> The development of the reading scale required some improvisation. The ETS proposal assumed that a block of reading items would include about twelve scalable reading exercises which would span a wide range of difficulty so that few students would be able to answer either none or all of the exercises correctly. Because many students would be given two or three reading blocks, there would be a large, random, subsample of students who responded to around 24 items. Twenty-four items is approximately the number of exercises usually suggested for estimating individual performance using the maximum likelihood method. However, we did not know all of the properties of the reading exercises at the time of proposal writing and, within the transition time constraint, we were not able to form blocks of exercises that met the "twelve exercises to a block with varying difficulty" criterion. Some of the blocks had fewer exercises, some students did not respond to all the exercises offered, many students were able to answer all exercises correctly, and many others scored less well than would be expected by chance. The total effect of these factors was that maximum likelihood estimates of reading proficiency were attainable for only a non-random subsample of NAEP students. (Beaton, 1987, p. 230)

[10] See also http://nces.ed.gov/nationsreportcard/ltt/performance-levels.aspx.

[11] These performance levels are not to be confused with main-NAEP achievement levels (*Basic, Proficient,* and *Advanced*). Achievement levels, established by a different process to define goals and expectations for student achievement, are important in the reporting and interpretation of main-NAEP assessment results. They are defined only for main-NAEP assessments and have never been established for LTT assessments. The earlier performance levels can still be used as an aid to interpreting results from LTT assessments.

Thus, the original analysis plan, which envisioned joint estimation of person and item parameters using LOGIST,[12] was abandoned. Mislevy (1985), recognizing that there was actually no need to estimate individual students' scores, took the lead in formulating a marginal maximum likelihood (MML) analysis strategy using BILOG (Mislevy and Stocking, 1989), with conditioning to "borrow strength" across responses from multiple examinees and to avoid bias in parameter estimates for subgroups. The new analysis strategy also required multiple imputation (i.e., sampling of "plausible values" from posterior distributions of ability for each examinee) to enable accurate estimation of population variances and of precision (Beaton, 1987). These procedures, initially presented in ETS technical reports, were more fully documented in subsequent publications (e.g., Mislevy, 1991; Mislevy, Beaton, Kaplan, and Sheehan, 1992; Mislevy, Johnson, and Muraki, 1992).

At the same time that ETS confronted the challenges of shaping NAEP's future, it also confronted challenges from NAEP's past. As the new NAEP contractor, ETS was committed to maintaining links to earlier NAEP administrations.[13] What ETS staff had to work with were raw data files and reports from administrations of objectives-based pools of exercises that had been written and assembled with little or no pilot or field testing and with no attention to questions of dimensionality (Cronbach, 2004; Lehmann, 2004; Johnson, 1988; Mislevy and Sheehan, 1987). Moreover, during previous NAEP data collections in some subjects, tape recordings of instructions and exercises had been used to guide students through the testing session. Students were expected to listen to the tape and follow along with the identical text in their exercise booklets. This procedure, referred to as audio-paced tape administration, was intended to assure that students did not spend too much time on any one exercise, had time to attempt all the exercises in the booklet, and were not unduly penalized for possible reading difficulties that were irrelevant to the content being assessed. However, a consequence of paced tape administration was that all students in a given testing session had to be given the same test form. Thus, under that original NAEP data collection design, referred to as "multiple matrix sampling," there was almost no overlap among exercises in separate booklets, and so the earlier NAEP data largely took the form of separate, nonoverlapping sets of exercises administered to distinct (although arguably randomly equivalent) samples of students.

Linkage to earlier assessments was accomplished via IRT, but again, some compromises were necessary. These analyses are described in detail in NAEP technical reports for reading (Mislevy and Sheehan, 1987) and mathematics (Johnson, 1988). For mathematics, IRT scaling extended only as far back as the second mathematics assessment, in 1977-78. By fitting lines to logit-transformed percents correct (p-values) for exercises administered in both 1972-73 and 1977-78 at each age level, the IRT-based scales were then extrapolated back to the first mathematics assessment. In addition, for both subject areas, only a subset of exercises could be used. This means that the NAEP LTT scales created at that time may not have been

[12] LOGIST and BILOG are two early computer programs for IRT estimation. For a detailed comparison of the two, see Mislevy and Stocking (1989).

[13] Initially, LTT trend lines were established for the areas of science, reading, mathematics, citizenship/social studies, and writing. For citizenship/social studies, one assessment prior to the redesign, in 1975-76, was linked to ETS-fielded NAEP assessments in citizenship/social studies (1981-82) and civics (1988), at which point this trend line ended. For writing, the 1969-70 assessment could not be linked successfully to the second assessment in 1984, and so the reported LTT trend in writing began with 1984 and ended with 1996 assessments (Jones and Olkin, 2004, pp. 562-563). The LTT trend in science continued until 1999. After careful consideration, the Governing Board chose not to include science in the next LTT assessment, in 2004, because item content was seriously out of date and it was unclear how comparable replacement content could be created (Governing Board, 2002; Stedman, 2009, p. 6). Thus, the focus of this white paper is solely on the LTTs in reading and in mathematics.

fully representative of the 1981-82 mathematics and 1979-80 reading objectives. A thorough analysis of relationships among the objectives, the full ECS item pools, and the item subsets actually scaled would help to inform the question of which domains of knowledge and skills the LTT assessments actually address, as discussed later in this paper.

To maintain trend lines, ETS administered two separate assessments. BIB-spiraling was introduced for the larger data collection, which was referred to as "main NAEP." The smaller data collection replicated earlier, audio-paced tape administration procedures with multiple matrix sampling, initiating what came to be called the NAEP Long-Term Trend component. LTT assessments followed the previous ECS administration schedule, with 13-year-olds assessed in the fall, 9-year-olds in the winter, and 17-year-olds in the spring. The first separate NAEP and LTT assessments under ETS occurred in reading in 1984 (1983-84 for the LTT) and in mathematics in 1986 (1985-86 for the LTT). This split is depicted schematically in Figure 1 by the bifurcation of one horizontal line into two, in 1984 for reading and in 1986 for mathematics. The figure offers a graphical representation of the dates of main-NAEP and LTT assessments. Assessments linked together as part of a common trend line are connected by horizontal lines, as are future assessments for which such linkages are anticipated. Places where trend lines have been affected by procedural changes are signaled by shifts from dashed to solid lines.

View Figure 1: Timelines for Long-Term Trend (LTT) and Main NAEP, showing past and future anticipated trend line changes. [Page 43]

Note that prior to 1990, the term "main NAEP" was used to refer to the larger, BIB-spiraled NAEP data collections in 1984, 1986, and 1988, in contrast to the "Long-Term Trend." Since 1990, however, the term "main NAEP" as applied to reading and mathematics assessments has been used almost exclusively to refer to data collections based on revised reading and mathematics frameworks introduced in 1990 and 1992, as discussed later in this paper. For the sake of clarity, therefore, the "main NAEP" assessments during the 1980s are referred to as "pre-1990 main NAEP" in the remainder of this paper.

NAEP and the LTT From 1986 to 1990

Pre-1990 main-NAEP and LTT assessments in both reading and mathematics were conducted by ETS for a second time in 1986. Further problems arose in 1986 in the pre-1990 main-NAEP reading assessment. The estimated performance of 17-year-olds declined sharply from 1984 to 1986, a drop so drastic that it was not credible (Beaton and Zwick, 1990). Performance of 9-year-olds also fell, although by a lesser amount, and performance of 13-year-olds increased slightly (Zwick, 1992b). This was the "NAEP reading anomaly," which led to a decision not to report any NAEP reading results for 1986. The reasons for these anomalous results may never be fully understood, but much of the problem was attributed to failures to account for changes in the locations where exercises appeared within booklets, as well as interactions among exercises (context effects). The reading anomaly was a wake-up call: Items (exercises) could not be treated as statistically independent indicators of achievement that would function in the same way regardless of context. IRT could not be relied on to produce comparable estimates unless care was taken to minimize changes to the administration contexts of individual exercises (e.g., antecedent exercises, placement within the overall testing session).

Numerous further changes were made over time. Pre-1990 main-NAEP assessments sometimes sampled both age- and grade-based student populations. Assessments were conducted at grades 4, 8, and 12 in 1984; then 3, 7, and 11 in 1986; and then 4, 8, and 12 in 1988 and thereafter. Age groups were defined by

birth dates between October 1 and September 30 for 17-year-olds and between January 1 and December 31 for 9-year-olds and 13-year-olds. NAEP under ECS had used separate testing windows for 9-, 13-, and 17-year-olds spaced throughout the school year. For main NAEP, these were shifted to a common testing window.[14] Sources of information for identifying students according to race/ethnicity changed. The subcontractor for NAEP sampling changed from Research Triangle Institute (RTI) to Westat, after which procedures for post-stratification adjustments were modified slightly. Technical improvements were also introduced, including use of the Partial Credit Model (Masters, 1982) for IRT calibration of some exercises and conditioning on principal components of background variables instead of the original variables themselves.

In addition, the plan set forth in the ETS proposal had called for BIB-spiral designs in which exercise blocks from two or more subject areas would appear together in some exercise booklets, enabling the estimation of student-level correlations among scores across subject areas. Unfortunately, even with the new MML/multiple imputation analytical strategy, assessing individual students in more than one subject area reduced the number of exercises administered within a single subject area to a level that was suboptimal. Thus, the originally proposed BIB design was replaced in 1988 with a "focused BIB" design, in which individual students were assessed in only a single content area (Beaton, 1990b).

Age-based versus grade-based populations are not directly comparable, which is one reason why NAEP reports are deliberately structured so as to discourage direct comparison between main-NAEP and LTT trend lines.[15] Thus, trend reports and cross-sectional reports generally appeared in separate publications. Nonetheless, up through the 1988 assessments, there was just one NAEP scale for reading and one for mathematics.[16] *The Reading Report Card, 1971-88: Trends From the Nation's Report Card* (Mullis and Jenkins, 1990) reports only a single trend line for each age/demographic group, with no differentiation between pre-1990 main-NAEP and long-term trends. Likewise, *The Mathematics Report Card: Are We Measuring Up? Trends in Achievement Based on the 1986 National Assessment* (Dossey, Mullis, Lindquist, and Chambers, 1988) reports only single trend lines, again with no differentiation between pre-1990 main-NAEP and long-term trends.

Main NAEP and the LTT From 1990 to 2012

Beginning around 1990, the NAEP governance structure changed significantly. As part of the 1988 reauthorization of the Elementary and Secondary Education Act of 1965 (P.L. 100–297, the Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988), Congress created the independent, nonpartisan National Assessment Governing Board (Governing Board). Although the National Assessment was placed within the National Center for Education Statistics (NCES) under the supervision of the Commissioner of Education Statistics, the Governing Board was given responsibility for formulating policy guidelines for the National Assessment, selecting subject areas to be assessed, developing assessment objectives and test specifications, and other matters. Notably, the Governing Board was also given the responsibility to "identify feasible achievement goals for each age and grade in each subject area to be tested under the National Assessment." In response to this mandate to

[14] These changes over time for main NAEP are detailed on the NCES website. See https://nces.ed.gov/nationsreportcard/tdw/overview.

[15] One of the few published direct comparisons of main-NAEP versus LTT trend lines was reported by Beaton and Chromy (2010), under the auspices of the NAEP Validity Studies Panel. See also Pellegrino, Jones, and Mitchell (1999, pp. 73-77).

[16] Discussion of separate IRT calibrations for subscales within each subject area is beyond the scope of this paper.

develop achievement goals, the Governing Board established a system of achievement levels (*Basic, Proficient,* and *Advanced*) for all subject areas and grade levels included in main-NAEP assessments (Vinovskis, 1998).[17]

Subsequent legislation affecting NAEP and the Governing Board included P.L. 107-279, the Education Sciences Reform Act of 2002. Over time, the size of the Governing Board was increased and changes were made to the terms for board members and the process by which new members are chosen, as well as language defining the responsibilities of the Governing Board, NCES, and the Commissioner of Education Statistics. In particular, the LTT assessments were called out in the 2002 legislation, which states in section 302(e)(1)(F) that the Governing Board (referred to in the legislation as the Assessment Board) shall, "consistent with section 303, measure student academic achievement in grades 4, 8, and 12 in the authorized academic subjects." It goes on to specify in sections 303(a) and 303(b)(2)(F) that "The Commissioner for Education Statistics shall, with the advice of the Assessment Board established under section 302, … 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." Thus, although the Governing Board has the authority to determine the schedule for NAEP assessments, including the LTT, any Board recommendation that could affect that law would depend on the commissioner's concurrence, and the LTT could not be discontinued altogether unless the law itself were changed or superseded.

Consistent with its legislative mandate, beginning around 1990, the newly formed Governing Board introduced new frameworks for main-NAEP assessments, developed through a broadly consultative process. As a result, the mathematics assessments were dramatically revised in 1990 and then significantly expanded in 1992. The reading assessment was dramatically revised in 1992. These historic changes were summarized in the 1992 NAEP technical report as follows:

> **Reading**: For the national assessment, a newly developed reading assessment was administered at grades 4, 8, and 12. This assessment was designed around questions requiring in-depth analysis of authentic, naturally occurring reading materials. A mixture of multiple-choice, short constructed-response, and extended constructed-response questions made up the survey; in aggregate well over half of the student assessment time was spent answering constructed-response rather than multiple-choice questions. …
>
> **Mathematics**: For the nation, the assessment that had been developed in 1990 was nearly doubled in scope for 1992 and administered at grades 4, 8, and 12. Assessment tasks included the use of four-function calculators at grade 4, scientific calculators at grades 8 and 12, and open-ended problem-solving questions at all grades. Manipulatives, rulers, and protractors also were available for use with portions of the assessment. Extended constructed-response questions were used on a wide scale for the first time in a NAEP mathematics assessment. In addition, estimation was assessed using audiotapes that paced students through the questions,[18] and complex problem-solving skills were assessed for the nation in special study blocks. ... (Campbell, Lazer, and Mullis, 1994, pp. 33-34)

[17] NAEP achievement levels are not to be confused with the performance levels, discussed earlier, which are still used as an aid to interpretation of LTT assessment results.

[18] Use of audiotapes to pace students through questions calling for estimation is intended to encourage students to work quickly, discouraging them from taking time for exact calculations, which would defeat the purpose of the estimation exercise. This format had been used previously, and so, technically, was not a change in procedures.

These changes were so substantial that new IRT scales were developed and new trend lines started. Thus, since 1992, there have been separate score scales for main NAEP versus the LTT. Long-term trends in reading are reported on the same scale for 9-year-olds, 13-year-olds, and 17-year-olds. The scale has a nominal range from 0 to 500, but virtually all scores fall between 100 and 400. Long-term trends in mathematics are likewise reported on a common 0-500 scale for all three age levels. These scales have been in place since before 1990. Main-NAEP reading results are also reported on a 0-500 scale, and main-NAEP mathematics results at grades 4 and 8 are similarly reported. Since 2005, main-NAEP mathematics results at grade 12 have been reported on a separate, 0-300 scale. The details of scale construction are complex and beyond the scope of this paper, typically beginning with separate IRT calibrations for two or more subscales within a subject area and grade level, after which subscale scores are combined to create a composite scale. Scores on this composite scale are then re-expressed on the 0-500 or 0-300 reporting scales. Even though scores for different age or grade levels are finally reported on a common scale, it may not be meaningful to make direct comparisons between score distributions for different grades (for main NAEP) or for different ages (for the LTT).

From 1990 to 2000, there was little further investment in the LTT component of NAEP, which continued essentially unchanged through assessments in 1994, 1996, and 1999.[19] By 2004, when the next LTT assessment was to occur, it was clear that revisions were necessary. To begin with, by law, testing accommodations for students with disabilities and for English-language learners had to be provided. Main NAEP had begun offering such accommodations in 1996, which not only satisfied a legal requirement, but also enabled the inclusion of a higher proportion of students than before, bringing the main-NAEP assessments closer to the ideal of assessing the full populations of students at each grade level.[20] LTT exclusion rates had crept upward through the 1990s (Allen, McClellan, and Stoeckel, 2005, pp. 17-18), and so a higher participation rate for LTT assessments was also seen as a benefit of the revised accommodation policy. Also in 2004, the LTT in writing was terminated by NCES because design weaknesses made the results unstable; and the Governing Board chose to suspend the LTT in science, primarily because the science content of many LTT exercises was seriously out of date.[21] Other significant changes were made to the reading and mathematics LTT assessments, including updates to reading passages, exercise formats, and contextual (background) questions and needed revisions to assessment administration procedures, including updates to allowable testing accommodations. Removal of science exercises also enabled reconfiguration of LTT exercise booklets so that each student was assessed in only one content area. (Previously, LTT exercise blocks from different subject areas had sometimes appeared together.) In addition, "I don't know" was dropped as a distractor choice for LTT exercises (bringing LTT into conformity with main NAEP) and paced-tape administration was ended. NCES designed a bridge study to address these and other concerns. A special bridge study was conducted to evaluate the effects of these changes on LTT performance.[22] Elimination of paced-tape administration

[19] Note that "essentially unchanged" does not mean that there were no changes. In 1994, for example, three out of 195 LTT reading items were treated as "new items" and recalibrated because they appeared to be functioning differently in 1994 than they had previously (Chang, Donoghue, Worthington, Wang, andand Freund, 1996, p. 365).
[20] Private-school students are included in some but not all NAEP assessments. How private-school populations are framed and sampled, when they are included, and at what levels of aggregation their results are reported are all issues beyond the scope of this paper.
[21] The Board called for technical studies to investigate the feasibility of continuing the LTT science assessment at some point in the future (Governing Board, 2002; Stedman, 2009, p. 6).
[22] The 2004 LTT bridge study included assessment of some students following prior LTT assessment procedures as closely as possible. For this purpose, science exercise blocks were included in booklets along with LTT reading and

and reconfiguration of exercise booklets in the 2004 LTT assessments also made it possible to field test new LTT replacement items for the first time in two decades. Note that development of replacement items was guided by the NAEP mathematics objectives from 1981-82 and the NAEP reading objectives from 1979-80, as well as the items included in the LTT prior to 2004.

LTT data collections occurred again in 2008 and 2012 following the 2004 revised procedures, with no significant further modifications. These assessments used some new LTT exercise blocks field-tested and calibrated in 2004. This made it possible to release some blocks of LTT items, which are now publicly available via the NAEP Questions Tool at http://nces.ed.gov/NationsReportCard/nqt.

Main NAEP continued to evolve, with updated reading assessment frameworks in 2002 and 2009 and with updated mathematics assessment frameworks for 2005 at grades 4 and 8 and for 2009 at grade 12. For 2005, a new grade 12 mathematics framework was developed. There were minor revisions at other times to the reading and mathematics frameworks. Main-NAEP trend lines were maintained through these changes with the exception that the grade 12 mathematics trend line was ended and a new trend line was started in 2005.[23] These and other changes affecting trend lines are more fully described by Beaton and Chromy (2010, Ch. 3).

Various parts of NAEP's rich history are further documented elsewhere (e.g., Beaton and Chromy, 2010; Bourque and Byrd, 2000; Haertel and Mullis, 1996; Jones, 1996; Jones and Olkin, 2004; Mullis, 1992; Perie, Moran, and Lutkus, 2005; Stedman, 2009; Vinovskis, 1998; Zwick, 1992a, 1992b).[24]

A Note on Bridge Studies

When NAEP population definitions, frameworks, administration conditions, or in some cases, analysis procedures change, "bridge studies" are typically conducted to determine whether the change has a material effect on assessment results. The precise design of each bridge study depends on the particular change being investigated, but the idea is to conduct an assessment twice at the same time, following both the old and the new procedures. In that way, any differences observed can be attributed to the procedural change itself (after accounting for inevitable statistical variation). Typically, the assessment following the old procedures is conducted on a reduced scale (i.e., with smaller student samples).

Ideally, the bridge study will demonstrate that the effect of the procedural change is small enough to ignore. In that case, trend lines can be maintained without interruption. In the event that the bridge study

---

mathematics items so as to avoid altering the context in which these reading and mathematics exercises were presented. The bridge study also examined the effects of new testing accommodations. Main-NAEP assessments in 1996 and 2000 had included subsamples assessed both with and without updated testing accommodations, and a similar, parallel data collection with and without updated accommodations was part of the 2004 LTT bridge study.

[23] As reported on the NCES website (https://nces.ed.gov/nationsreportcard/mathematics/frameworkcomparison.aspx), the 2009 revisions to the grade 12 mathematics framework "added objectives addressing mathematics content beyond what is usually taught in a standard 3-year mathematics course in high school … [so as to] help NAEP report how well prepared twelfth-grade students are for post-secondary education and training." A bridge study confirmed that, despite these framework revisions, grade 12 mathematics results from 2005 and 2009 were directly comparable, and so the new grade 12 mathematics trend line was continued from 2005 to 2009 without further interruption.

[24] Informative summaries also appear on the NCES website at http://nces.ed.gov/nationsreportcard/about/newnaephistory.aspx and http://nces.ed.gov/nationsreportcard/about/naephistory.aspx.

shows a material effect of some change, assessment results from old and new procedures can be compared to estimate the size of that effect. Again depending on the nature of the change, the interruption might or might not require defining a new set of NAEP reporting scales. Beaton explained bridge studies as follows, with reference to the original split in 1984 between main NAEP and the LTT:

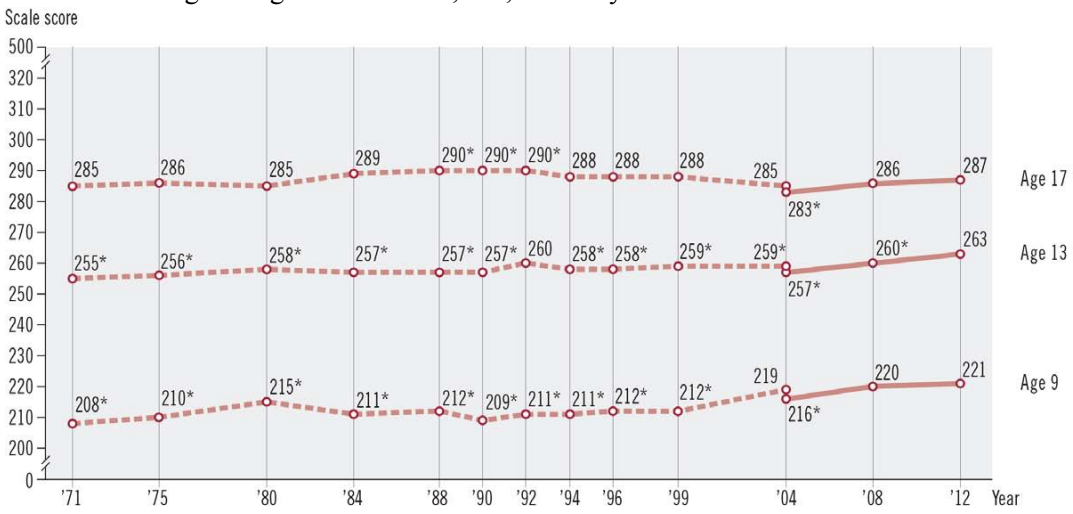
> It is useful at this point to consider the consequences of introducing a new design into an existing measurement system. There is a clear tension between the need to maintain constant measurement procedures in order to estimate changes in performance and the desire to continue to improve the assessment by using the most modern, best available technology. The new design introduced in 1984 responded to this tension by assessing student achievement in two ways: in one set of samples using the methods of past assessments and in another set using the best available methodology. The samples using the methods of the past were called "bridge" samples, since they provided bridges to the performance of students in past assessments. The result was parallel assessments, using different technologies, that could be compared and for some purposes, perhaps, equated. In this way, innovations could be introduced without losing comparability with the past. Although this flexibility to introduce innovations while maintaining trends has come at the cost of increased complexity, the flexibility does allow NAEP to be responsive to the information needs of policy makers while maintaining the scientific requirements of sophisticated survey research. (Beaton, 1990b, p. 5)

Even if some procedural innovation effects a material change, trend lines can often be maintained by simply incorporating the estimated effect of the innovation via some adjustment and continuing to report new observations on the old trend line. This has been the preferred approach. Whenever possible, such minor adjustments are simply absorbed into the scale transformations used to express IRT scale scores on NAEP reporting scales. A concern with this approach, however, is that the bridge study samples tested following old procedures are typically smaller, sometimes much smaller, than the samples for regular NAEP assessments. Consequently, bridge study estimates of the magnitudes of effects of procedural changes are somewhat imprecise, and so reliance on these estimates affects trend line accuracy.[25]

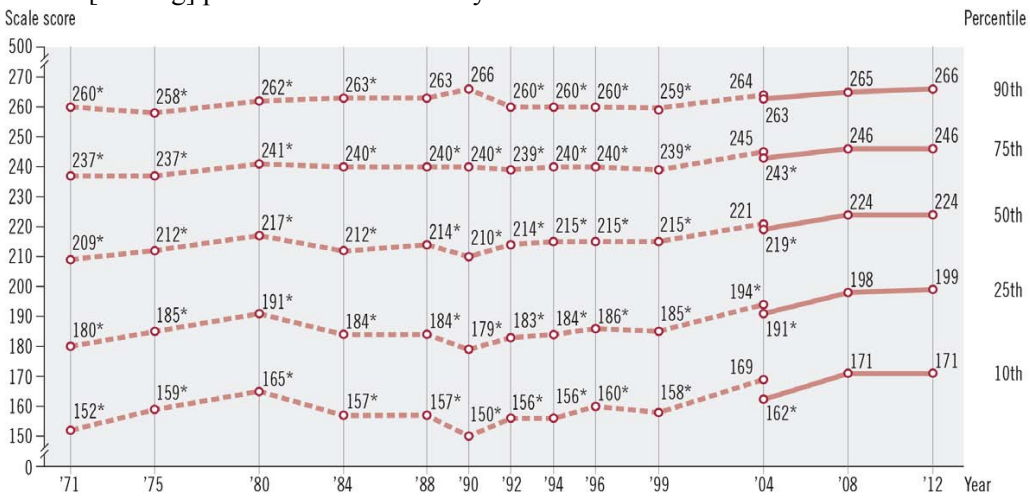
Unfortunately, the effects of changes in assessment frameworks, populations, or procedures cannot always be accounted for by scale adjustments. It may be difficult or impossible to maintain unbroken trend lines if the effects of a procedural change are markedly different for different parts of the tested population or if the definition of the assessed population itself is changed. This can be seen in the following two graphics taken from the 2012 NAEP LTT report (NCES, 2013, pp. 7-8):

[25] Comparisons of performance across greater spans of time will tend to be less accurate than those close in time, in part because random errors introduced by any intervening bridge study adjustments are cumulative, but also because there is some small, random error in linkages from one assessment to the next, even when procedures are unchanged. These year-to-year uncertainties will gradually accumulate over time.

Trend in NAEP reading average scores for 9-, 13-, and 17-year-old students



Trend in NAEP [reading] percentile scores for 9-year-old students



Note the breaks in trend lines shown at 2004. In the first of these graphics, it is seen that the average, overall effect for 9-year-olds of the various 2004 procedural changes to the LTT was to reduce the mean scale score in reading by 3 points, from 219 to 216. The asterisk indicates that this change was statistically significant. But the second, more detailed figure for only 9-year-old reading shows that the effect of the procedural changes was not a simple 3-point change everywhere in the distribution. At the 90th percentile, the effect was only 1 point (from 264 to 263), whereas at the 10th percentile, the effect was 7 points (from 169 to 162). The 2004 changes included new testing accommodations, so that under the new procedures, fewer students with disabilities were excluded. This subtly changed the definition of the tested population, in a way that affected primarily the lower part of the score distribution. As shown in the two graphics, when procedural effects prevent the direct comparison of old and new results, then trend lines must be restarted. In some such cases, new scales must be created. Both sets of bridge study results (i.e., old-procedure results and new-procedure results) are reported for the same year, each on its own trend line. Prior assessments can then be compared directly with bridge study results under old procedures, and future assessments can be compared directly with bridge study results under new procedures.

Future Plans for Main-NAEP Assessments in Reading and Mathematics

Preparations have been underway for some time for main NAEP to transition from paper-and-pencil to digitally based assessment. A new framework was adopted for writing, starting with the 2011 assessment at grades 8 and 12 and with the 2017 assessment at grade 4. This framework is computer-based, and so main-NAEP writing assessments at grades 8 and 12 have been computer-based since 2011. Current plans call for 2017 main-NAEP assessments in reading and mathematics, as well as writing at grades 4 and 8, to be conducted using tablets with keyboards.[26] A bridging study in 2017 will include smaller samples tested concurrently with paper-and-pencil test forms instead of tablets, so that administration mode effects can be monitored. The expectation is that mode effects will be accounted for in the scaling of digital exercises, and therefore, it will be possible to continue main-NAEP trend lines unbroken, pursuant to a Governing Board resolution adopted in May 2015.[27]

Summary

The Nation's Report Card offers two histories of achievement trends in reading and in mathematics, reported on separate, though similar, scales; for distinct, though similar, populations; covering different, though overlapping, time periods. LTT assessments from the early 1970s to 2012 are reported for schoolchildren at ages 9, 13, and 17 on LTT score scales; and main-NAEP assessments from the early 1990s to the present are reported for schoolchildren at grades 4, 8, and 12 on more recent score scales. There are also main-NAEP assessments from before 1990, reported for grade-based samples and sometimes for age-based samples on the earlier score scales still used for the LTTs.

Writing about the NAEP reading anomaly, Albert E. Beaton admonished, "*When measuring change, do not change the measure*" (Beaton, 1990a, p. 10, italics in original).[28] The narrative of the LTT has been that it has honored this admonition, serving as a stable anchor, measuring achievement in the same way, with the same frameworks, and mostly the same items since NAEP began. This account contrasts the LTT with main NAEP, which has been more adaptive in response to shifts over time in curriculum and instruction and to changing policy concerns, as well as new measurement innovations. The report on *NAEP 2004 Trends in Academic Progress* stated that "the long-term trend instruments do not evolve based on changes in curricula or in educational practices; in this way, the long-term trend assessments differ from the main national and state NAEP assessments" (Perie, Moran, and Lutkus, 2005, p. 91). In a paper commissioned for the 20th anniversary of the Governing Board, a former Governing Board member asserted in passing that "the LTT … consists of the same test questions in reading, mathematics, and science that were first offered in the 1970s" (Ravitch, 2009, p. 5). In another paper commissioned for that same anniversary event, Stedman further documented such assertions:

> The strongest argument for retaining the long-term trend assessment is that it has tested the same items for more than three decades and so has a unique ability to track changes. It is supposedly based on a single curricular framework and unchanging tests first administered in 1969 or the early 1970s. This is a widely held belief. Secretary Spellings described the assessment as “using

[26] The 2017 grade 12 writing assessment was cut from the schedule due to budget constraints. Digitally based writing assessments at all three grade levels are scheduled for 2021.

[27] See https://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-trend-and-dba.pdf.

[28] As noted in an earlier footnote, it is not quite accurate to refer to the comparison described as a measure of "change," because two distinct samples, representing two distinct populations, are being compared. Nonetheless, the shorthand reference to "change" should cause little difficulty so long as the actual nature of the comparison is not forgotten.

> the same exact test in reading and mathematics for over 30 years" (U.S. Department of Education, 2005). NCTM (2004) noted, "The same test items have been used for mathematics since 1973." Even NCES … perpetuates this view, stating that content "has remained essentially unchanged since first administration (1971 for reading, 1973 for mathematics), although some changes were initiated in 2004."[29] (Stedman, 2009, p. 28)

As this brief review has shown, however, the reality is that both main NAEP and the LTT have changed over time. In the sentence following his admonition to "… *not change the measure*," Beaton (1990a, p. 10) continued, "Precise implementation of this dictum is, of course, impossible in actual practice. In fact, NAEP has modified its measurement instruments by rearranging and reformatting assessment exercises since it began measuring trends." Zwick (1992b, p. 206) likewise acknowledged that, "Like any long-term study of trends in educational achievement, NAEP is subject to competing pressures … On one hand, the measurement of changes in performance is facilitated by the retention of existing assessment instruments, administration procedures, and analysis techniques. On the other hand, requiring the assessment to remain identical over time would prevent the introduction of new curriculum concepts and measurement technology."

Stedman (2009) recounted some of the same history as set forth in this paper, amply attesting to significant changes in LTT content and administration procedures prior to 1990. In an appendix, by way of illustration, he provided the numbers of linking items (exercises) used at various times to maintain trend lines for 17-year-olds. For reading, these links were based on 71 items in 1980, 53 items in 1984, and 87 items in 1988. For mathematics, among other details, Stedman reported that just 61 items were used to link the 1978 and 1982 assessments for 17-year-olds. Total numbers of items (i.e., linking items plus additional items administered and calibrated along with linking items) also varied widely across LTT assessments, both in reading and in mathematics.

Throughout the 1990s, the LTT was in fact continued virtually unchanged, but this effort to follow Beaton's dictum not to change the measure, using the same test forms for years, resulted in cumulative obsolescence, which ironically may have impaired NAEP's ability to measure change. The meanings of fixed items can and do change over time. This is abundantly clear in science, but it is seen to a greater or lesser extent in all subject areas. Today, all NAEP assessments, including the LTT since 2004, are designed in a way that permits retiring old, obsolete items and replacing them with fresh items that measure the intended framework or objectives.

All this is not to deny that the LTT has measured the same content for decades. Since the mid-1980s, the LTT has continued to represent content defined by the NAEP mathematics objectives from 1981-82 and the NAEP reading objectives from 1979-80, subject to some curtailment because not all early items could be scaled. Even when the LTT was updated in 2004, an effort was made to "reverse-engineer" the objectives underlying the existing LTT item pools so that new LTT items could be added without

[29] The quotation Stedman provides from the NCES website, at http://nces.ed.gov/nationsreportcard/about/ltt_main_diff.aspx, has since been updated, and now correctly explains that content "has remained relatively unchanged since 1990. In the 1970s and '80s, the assessments changed to reflect changes in curriculum in the nation's schools. Continuity of assessment content was sufficient not to require a break in trends." Likewise, as compared with the *NAEP 2004 Trends* report, the language in the report on *NAEP 2012 Trends in Academic Progress* (NCES, 2013) is more cautious, and more accurate, in describing the continuity of the LTT prior to 1990.

changing the constructs assessed.[30] And, although reading and mathematics objectives evolved during the 1970s, the LTT does provide some linkage all the way back to the very first NAEP reading and mathematics assessments in the early 1970s. It does not, however, continue to reflect the curriculum content or the educational priorities of NAEP's very beginning.

Further significant changes loom as main NAEP continues the transition to a digital platform in 2017. These changes might be taken to imply that the LTT is increasingly irrelevant, or that the LTT is more important than ever. Regardless, the LTT assessment for 2016 has been twice postponed, first to 2020 and then to 2024. The LTT stands at a critical juncture. Its future is unclear.

What Do the LTT Assessments Measure?

Before turning to options for the LTT's future, it may be helpful to review publicly available information as to just what it is that the LTT assessments in reading and mathematics actually assess. This question turns out to be surprisingly difficult to answer. Recall that there are no fully developed content frameworks for the LTTs. Rather, the LTTs began as collections of exercises, operationalizing lists of objectives that changed over time. Some fraction of those exercises survived being released (i.e., were kept secure for reuse) and also survived screening on technical criteria, screening for bias, screening for outdated or obsolete content, or elimination on any other grounds. These surviving exercises, sometimes revised, augmented with some additional exercises intended to measure the same content, became the LTT exercise pools.

The NCES website (https://nces.ed.gov/nationsreportcard/ltt/moreabout.aspx) offers the following description of what is measured by the LTTs in mathematics and reading, and much the same language appears in recent NAEP trend reports:[31]

> **Mathematics:** The long-term trend mathematics assessment required students to respond to a variety of age-appropriate questions. The assessment was designed to measure students'
>
> - knowledge of mathematical facts,
> - ability to carry out computations using paper and pencil,
> - knowledge of basic formulas such as those applied in geometric settings, and
> - ability to apply mathematics to daily-living skills such as those involving time and money.
>
> …
>
> **Reading:** The NAEP long-term trend reading assessment measures students' reading comprehension skills using an array of passages that vary by text types and length. The assessment was designed to measure students' ability to
>
> - locate specific information in the text provided,
> - make inferences across a passage to provide an explanation, and

[30] This "reverse-engineering" may be described in internal contractor documents, but there are no LTT frameworks comparable in scope and organization to NAGB-developed frameworks for main NAEP. Formally developed and adopted frameworks and specifications for the LTT would be enormously helpful in systematizing the ongoing creation of replacement items, especially if the LTT transitions to a digital platform.

[31] Stedman (2009, pp. 29-30) also summarizes several brief descriptions of LTT content from earlier NCES reports, some of which appear inconsistent with one another.

- identify the main idea in the text.

Some released LTT exercises are available on the NCES website, formatted as a "Sample Questions Booklet."[32] These include both multiple-choice and constructed-response exercises for both reading and mathematics. The 2011-12 "Sample Questions Booklet" also includes more detailed descriptions of LTT content for the 2011-12 assessments in both mathematics and reading at all three age levels assessed (included as Appendix A to this paper). The descriptions of the five content topics in mathematics paraphrase parts of the "content outline" included as Appendix C in the *1981-82 NAEP Mathematics Objectives* (NAEP, 1981a, pp. 33-35), although language is updated, some subtopics have been removed and a few added, and a sixth category from the *1981-82 Mathematics Objectives* ("technology") has been omitted. Likewise, the descriptions of four mathematics "process domains" largely quote or paraphrase the first four of the five "process" categories from the 1981-82 objectives (NAEP, 1981a, pp. 14-16), with the fifth category ("attitudes toward mathematics") being omitted. The "target percentages" by content topic for each age level differ from those presented in the 1981 document.

The description of the LTT in reading in the 2011-12 "Sample Questions Booklet" may be based in part on the *Reading Report Card* reporting trends in reading from 1971-84 (NAEP, 1985), although it includes considerable detail not found in that trends report. As with mathematics, the reading description appears to have been updated slightly from similar descriptions published in sample questions booklets from the 2003-04 and from 2007-08 LTT assessments. The classification of exercises into categories of "expository, ""narrative," and "document and other" was not located in any previous NAEP reading objectives documents, and no percentages of items by text type appear in the sample questions booklets from earlier assessments.

LTT content was examined in a recent study by Dickinson, Taylor, Koger, Moody, Deatz, and Koger (2006). These researchers documented the alignment between (1) the 2004 LTT exercises in reading and mathematics at ages 9 and 13 and (2) the 2003 main-NAEP assessment frameworks[33] and exercises in reading and mathematics at grades 4 and 8. Dickinson and her colleagues lamented the absence of any LTT content frameworks that they would otherwise have compared with main-NAEP frameworks, but they were able to report that, overall, they rated both main NAEP and LTT items as being of high quality. Referring to Webb's (1997, 1999) system for coding "depth of knowledge," they also reported that "it is clear that Main NAEP items tend to assess higher depth of knowledge levels than LTT items. In other words, students must use complex processing more often to answer items for Main NAEP than for LTT items" (Dickinson, et al., 2006, p. 21). In the end, these authors concluded that "the LTT and Main NAEP assessments measure specific content objectives that are different, and therefore the two assessments cannot be considered interchangeable. … Main NAEP targets content that is not being measured by LTT test items. [Conversely], though LTT items can be linked to general-level content strands within the Main NAEP frameworks, … if a subset of Main NAEP items were selected in an attempt to replace the LTT item pool, matching the content intentions of the LTT assessment would be difficult if not impossible given the lack of LTT content frameworks" (Dickinson, et al., 2006, p. 24).

---

[32] See http://nces.ed.gov/nationsreportcard/subject/_commonobjects/pdf/demo_booklet/2011_12_sqb_ltt.pdf. Individual released exercises from 2004 and 2008 may also be viewed using the NAEP Questions Tool. To view these exercises, beginning at http://nces.ed.gov/NationsReportCard/nqt, choose "Search Questions," then select either "LTT Mathematics" or "LTT Reading" in the "Select Subject" pull-down menu and proceed accordingly.

[33] Note that these main-NAEP frameworks have since been revised.

So, accepting the summary descriptions on the NCES website and in NAEP reports, as well as the summaries and the released exercises themselves from the sample exercise booklet accessible online, we may conclude that the content and processes assessed by LTT exercises fall within the range of typical curricular expectations for grades 4, 8, and 12, but that more advanced topics and more complex processes at these grade levels are largely omitted. It is likely that many LTT exercises at ages 9, 13, and 17 are better aligned with contemporary curriculum and instruction at grades 3, 7, and 11, respectively, than grades 4, 8, and 12. Based on inspection of the released exercises in the "Sample Questions Booklet," none of the released exercises available appears to be out of date, nor are there any that appear likely to be judged by thoughtful and knowledgeable reviewers as inappropriate or as addressing content that should not be taught. There would probably be consensus that the LTT omits many important topics, although it might be difficult to reach agreement as to exactly what those important omitted topics were. Obviously, these matters could best be informed by systematic studies of the entire LTT exercise pool.

Arguments For and Against Preserving the LTT in Its Current Form Versus Dropping It Altogether

As already described, Title III of P.L. 107-279 (titled the "National Assessment of Educational Progress Authorization Act") requires that the Commissioner for Education Statistics "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." In addition to this statutory requirement, most arguments for maintaining separate LTT assessments are premised on one or both of the following propositions:

- LTT assessments are unique in tracking achievement over a very long time span.
- LTT assessments measure knowledge and skills that are both important and distinct from what is measured by main-NAEP assessments.

With some qualifications, the historical review and content summary in this paper support both of these contentions. The LTT assessments might best be regarded as an anchor, assessing a fairly low-level, traditional subset of contemporary curricular objectives. For mathematics, comparison of the LTT content area distributions in Appendix A with the percentage distributions by grade and content area in the *2013 Main-NAEP Mathematics Framework* (Governing Board, 2012) confirms that the LTT places much greater weight on numbers and numeration, especially at ages 13 and 17, than the main-NAEP framework places on "number properties and operations," especially at grades 8 and 12. For reading, comparison of the LTT descriptions in Appendix A with the *2015 Main-NAEP Reading Framework* (Governing Board, 2015) confirms that the LTT places more weight on expository texts at age 9 than does main-NAEP reading at grade 4. Also, at all three grade levels, the LTT relies on shorter reading passages than main NAEP (up to 250, 500, and 800 words for ages 9, 13, and 17 on the LTT versus up to 800, 1,000, and 1,500 words for grades 4, 8, and 12 for main NAEP, respectively). Although information is less readily available concerning differences in the skills and thinking processes elicited by main NAEP versus the LTT, comparison of the brief descriptions of LTT target skills on the NCES website versus current main-NAEP frameworks in reading and mathematics, as well as the depth-of-knowledge comparisons presented by Dickinson and colleagues (2006), suggests that main-NAEP exercises by and large call for more complex reasoning as well. For both content areas, then, the LTT appears to assess a subset of main-NAEP content and skills, although as noted by Dickinson and colleagues (2006), it might be difficult to delineate subsets of main-NAEP exercises that mirrored the LTT exercise pools.

A more specific argument in favor of continuing the LTT in its current form is the risk of jeopardizing long-term trends if the LTT were modified. This argument might carry more weight in light of impending

major changes, as main NAEP continues its transition to a digital platform. Keeping the LTT component unchanged might be viewed as an "insurance policy" in the event of unforeseen problems akin to the 1986 reading anomaly, although a counterargument would be that, with the next LTT now scheduled for 2024, the LTT would be of little help in any case if main-NAEP trends were compromised in 2017.

It can also be argued that, in addition to being required by law, the LTT's age-based achievement trends offer a valuable counterpoint to the grade-based trends from main NAEP. Contrasts between age-based and grade-based gaps and trends might inform contemporary policy questions arising from the societal trend toward children starting formal schooling at a later chronological age (Deming and Dynarski, 2008) as well as differential rates of grade retention for different racial/ethnic groups (Nagaoka and Roderick, 2004). Performance trends over time for 13-year-olds (rather than eighth graders) might also serve as a better basis for comparison with performance trends on international assessments that also rely on age-based samples. Even apart from such specific policy questions, the fact that fourth graders today are, on average, older than fourth graders 10 or 20 years ago confounds the interpretation of trend lines for successive fourth-grade cohorts, and similarly for 8th and 12th graders. Age-based trend lines therefore offer important context for interpretations of grade-based trend lines.

Finally, as main NAEP evolves in response to changing curricular priorities and expectations for schooling outcomes, both proponents and critics of reforms such as the Common Core State Standards may wish to document achievement trends brought forward from an earlier, perhaps simpler, time, to see whether more "basic skills" are compromised by increasing emphasis on more complex learning objectives. One of the original architects of NAEP, Lee J. Cronbach (1963), argued for the importance of testing potentially valued instructional objectives *not* emphasized in a given curriculum. The same argument can be elevated from the level of course evaluation to the level of achievement profiles for the nation's youth:

> An ideal evaluation would include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which this curriculum directs substantial attention. If you wish only to know how well a curriculum is achieving its objectives, you fit the test to the curriculum; but if you wish to know how well the curriculum is serving the national interest, you measure all outcomes that might be worth striving for. One of the new mathematics courses may disavow any attempt to teach numerical trigonometry, and indeed, might discard nearly all computational work. It is still perfectly reasonable to ask how well graduates of the course can compute and can solve right triangles. Even if the course developers went so far as to contend that computational skill is no proper objective of secondary instruction, they will encounter educators and laymen who do not share their view. If it can be shown that students who come through the new course are fairly proficient in computation despite the lack of direct teaching, the doubters will be reassured. If not, the evidence makes clear how much is being sacrificed. (Cronbach, 1963, p. 680)

The principal arguments against maintaining the LTT in its current form are that it is expensive, that maintaining two trends is confusing, that performance on outdated content is no longer of interest to policymakers or other stakeholders, and that a range of changes in schooling, in assessment technology, and in society at large are rendering it irrelevant and possibly invalid.

Cost concerns seem always to have weighed heavily on discussions of NAEP assessment schedules, including the decisions to postpone the next LTT data collection first to 2019-20 and then to 2023-24.

Resources are limited, and trade-offs among competing priorities are inevitable. It is difficult to pursue this argument any further here, because that would require delving into the specifics of costs for the LTT and for competing NAEP priorities and then weighing these against perceived benefits, all within the constraints of historical commitments and statutory requirements.

The argument that the LTT should be discontinued because having two trends is confusing seems weak. This argument was advanced in the National Research Council report, *Grading the Nation's Report Card* (Pellegrino, Jones, and Mitchell, 1999, p. 73), but as argued by Stedman (2009), disparate findings from LTT versus main-NAEP trends might equally well be cited as support for *maintaining* both assessment components, so as to better understand sources of differences. Beaton and Chromy (2010) found that for time periods where main-NAEP and LTT trend lines overlap, they do not entirely agree. After documenting procedural differences that complicated direct comparison of main-NAEP versus LTT findings, these authors explained their procedure for transforming reported results for the two assessment components into a common form. They then compared annual changes expressed as scale points per year for reading and for mathematics, comparing 9-year-olds assessed on the LTT with 4th graders assessed in main-NAEP and similarly for 13-year-olds and 8th graders and for 17-year-olds and 12th graders. They discussed whether annualized changes from main NAEP and from the LTT were statistically different from zero and also whether these changes for the two assessments were statistically different from one another. Although some differences between the two assessments appeared substantively meaningful, virtually none were statistically significant. This was unsurprising given the low power to detect differences (between assessments) in differences (over years, as captured by regression coefficients representing fitted trend lines).

A related argument holds that LTT trend lines are becoming redundant now that main-NAEP trend lines reach back more than 25 years. This argument also seems weak for three reasons. First, as shown in Figure 1, the LTT trend lines still cover a long and significant historical period not covered by main-NAEP trend lines. Although the historical LTT trend line would still be available if the LTT were discontinued, information about contemporary performance on LTT content would be lost, and direct comparisons with that historical period would no longer be possible. Second, LTT trend lines in mathematics have been interrupted, and when the reading framework was revised in 2009, it was an open question whether trend lines could be maintained. Future interruptions of main-NAEP trend lines are likely. Finally, main NAEP and the LTT assess distinct mixes of academic content and skills, reflecting distinct curricular priorities. The more stable, more conservative LTT trend lines provide a valuable counterpoint to the more adaptive main NAEP trend lines, which reflect more contemporary conceptions of valued schooling outcomes.

As to the idea that having two trend lines per se is simply confusing, a measurement specialist might counter that much mischief has arisen from the naive notion that "a test measures what it says at the top of the page" (Braun and Mislevy, 2005, p. 492). No one is well served by hiding evidence that not all reading tests, nor all mathematics tests, measure the same thing.

The argument that performance on outdated content is no longer of interest to policymakers or other stakeholders has already been addressed. Long-term trends will certainly not be of interest to everyone,

but as noted, both advocates and skeptics of recent curriculum reforms may wish to continue tracking achievement trends that reach back to a time before those reforms were initiated.[34]

Arguments that changes in schooling, technology, and society at large are rendering the LTT obsolete require more careful consideration. As the day-to-day activities of schooling evolve over the years, test tasks that once seemed familiar may come to seem foreign. Specific kinds of math problems that were once practiced by eighth graders but are now taught at earlier grade levels may have become more challenging when appearing on eighth-grade examinations. As children come to spend proportionately more time working in groups or applying mathematics and reading skills in the context of more complex, project-based instructional activities, or as assessment is better integrated into instructional activities and separate, stand-alone tests are used less often, the routines and expectations of the testing situation may become less familiar, and students' test performance may be affected. Indeed, "as a result … labels such as reading and writing do not necessarily retain the same meaning over time … and may not mean the same thing for bridge and main assessments that occur within a single year" (Zwick 1992b, p. 207).

Approaches to Integrating (or Blending) LTT and Main-NAEP Assessments

It would seem ideal, perhaps, to find some middle way between continuing the LTT unchanged and discontinuing it altogether. The goal would be to realize the benefits of maintaining the LTT while minimizing the costs. To summarize, in addition to the requirement in law that the LTT be continued, the principal benefits include maintaining long-term trends reaching back to the 1970s; reporting separately on still-relevant achievement outcomes dating from the days before the Common Core State Standards; and maintaining trends on achievement for populations defined by age, to complement main-NAEP assessments of populations defined by grade level. The costs include dollar costs; increased complexity of data collection, analysis, and reporting; and the testing burden.

Compromise positions that reject both continuing the LTT unchanged and dropping it altogether have been advanced repeatedly. In a unanimously adopted policy statement, the Governing Board called for a transition toward making main NAEP the primary source of information for trend reporting, while continuing LTT assessments on a less frequent basis (Governing Board, 1996, pp. 9-10). Since then, the NAEP program has in fact shifted in that direction. As shown in Figure 1, LTT assessments have become less frequent and main-NAEP trend lines have grown to cover longer and longer periods of time. The 1996 Governing Board statement also called for periodic updates to the LTT, with bridging studies to maintain trends, and an update of the sort envisioned was enacted a few years later, via the 2004 LTT bridge study.

A similar proposal was advanced in an NRC evaluation of NAEP, *Grading the Nation's Report Card*. That report called for "[reducing] the number of independent large-scale data collections [referred to in the report, using terminology current at the time, as national NAEP, state NAEP, and trend NAEP] while maintaining trend lines, periodically updating frameworks, and providing accurate national and state-level estimates of academic achievement" (Pellegrino, Jones, and Mitchell, 1999, p. 56).

[34] A reviewer raised the possibility of tracking performance on earlier content by analyzing a subset of main-NAEP items. As noted by Dickinson, et al. (2006), it is not clear how, or even if, a subset of main-NAEP items aligned with the objectives assessed by the LTT could be identified.

The technical challenges in fully merging main-NAEP and LTT assessments are daunting, however, perhaps even insurmountable. In addition to distinct exercise pools and the absence of well-defined frameworks for LTT assessment content and target skills, differences between LTT and main-NAEP assessments include sampling from populations defined according to age versus grade level; different testing windows (i.e., testing at different times during the school year); and incompatible exercise booklet formats (with three 15-minute versus two 25-minute blocks of cognitive exercises). Past experience, including the 1986 reading anomaly, has shown that even small changes in procedures can disrupt trend lines. Thus, simply merging LTT exercises into main-NAEP booklets would carry the risk of seriously disrupting both LTT and main-NAEP trend lines.

A general caution concerning changes to NAEP procedures appeared in a 2012 white paper on possible directions for a NAEP redesign. As part of its initiative on the future of NAEP, NCES convened a diverse group of technical experts for a summit meeting in August 2011 and a group of state and local stakeholders for a second summit meeting in January 2012. A panel of experts was then assembled to synthesize the presentations and discussions from these two meetings. That panel prepared a white paper in which they cautioned that even "seemingly innocuous changes in the underlying survey and psychometric models can take years to understand and validate, and more years before they become part of NAEP operations." Nearly all of that Expert Panel's recommendations were to initiate investigations of promising potential improvements to NAEP infrastructure and operations; few were for immediate changes (Expert Panel on the Future of NAEP, 2012, p. 9).[35]

In a background paper commissioned to inform work of the NRC committee that authored the 1999 report on *Grading the Nation's Report Card*, Kolen (2000) offered a thoughtful analysis of options concerning the LTT. These included continuing the LTT under explicit guidelines permitting small, periodic updates; eliminating the LTT (following linking studies to connect main-NAEP and LTT trend lines as best possible); and modifying main NAEP so that main-NAEP results could be reported on both LTT and main-NAEP scales. These were Designs 2, 3, and 4 in Kolen's paper. His Design 1 was the continuation of the separate LTT, unchanged. He concluded that the most conservative choice, maintaining a separate LTT unchanged, was safest in the sense of providing the greatest assurance that long-term trends could be continued, and also (probably) providing the best insurance against distortions to trend lines in the event that main NAEP began to drift toward high-stakes uses (a threat salient at the time of his writing, in the light of discussions around a possible "voluntary national test").[36] Many of the changes Kolen recommended in connection with his Design 2 (a separate LTT with periodic modest updates) were in fact carried out some years later in the 2004 LTT bridge study. Concerning his two remaining possibilities (either linking or integrating the two NAEP components), Kolen cautioned that:

> Both designs require conducting complex linking studies, making strong statistical assumptions, and being supported by an extensive research program for developing linking procedures that work in the NAEP context. The outcome of this research program is difficult to predict. Possibly, procedures could be developed that allow for linking assessments as different as long-term trend NAEP and main NAEP or as different as new and old main NAEP. However, it is also possible

[35] The author of this paper, Edward Haertel, also chaired the Expert Panel on the Future of NAEP.

[36] Note that, then as now, NAEP law prohibited any reporting of individual student or school results. Very briefly, the concern at the time was that if a "voluntary national test" were created, aligned with NAEP frameworks, and if individual scores from such a test were reported in metrics resembling NAEP scales, then the motivational context for future NAEP assessments might be altered.

> that the results of the research will indicate that changes in main NAEP assessments need to be much more tightly constrained than is presently the case. (Kolen, 2000, p. 148)

Kolen's analysis provides helpful background, but his paper did not address the fact that LTT versus main-NAEP data collections occur at different times in the school year. Because this appears to be a major challenge in combining the two assessments, and because both the LTT and main NAEP have changed significantly since Kolen was writing almost 20 years ago, the specific details of his proposals are not considered further in this paper.

The plan discussed below differs substantially from any of Kolen's (2000) options. It would preserve separate reporting of LTT and main-NAEP trend lines following a partial integration of main NAEP and the LTT. It would be expected to result in significant cost savings and to increase the long-term viability of the LTT. This plan would require bridging LTT trend lines (as happened in 2004) but would not be expected to have any material effect on main-NAEP procedures or trend lines. Note that in the following discussion, it is convenient to stay with the terminology of "booklets," even though, with the move to a digital platform, paper-and-pencil booklets may no longer be used. It is convenient to assume here that the main-NAEP BIB-spiral design will be maintained, with two exercise blocks (plus contextual questions, etc.) in the booklet administered to each examinee, and that the main-NAEP sampling design will be largely unchanged, although the options proposed here do not depend critically on either of these assumptions.

Testing Window and Assessment Platform

The 2004 LTT bridge study brought the LTT into closer correspondence with main NAEP, but substantial differences remain. LTT data collections still occur in the fall for 13-year-olds, in the winter for 9-year-olds, and in the spring for 17-year-olds, in contrast to the January-March main-NAEP data collection window at all three grade levels. The LTT still samples student populations defined by age, not grade. Most important, of course, the content assessed by the LTT differs from that assessed by main NAEP. In addition, contextual questions, used as conditioning variables in the complex NAEP scoring process, are different for main NAEP versus the LTT, although as noted by a reviewer, changes to conditioning variables need not disrupt trends, provided the specific variables used to define reporting subgroups remain unchanged. Another major divergence between main NAEP and the LTT looms, with the impending transition of main NAEP reading and mathematics assessments to a digital platform. Finally, as already noted in passing, NAEP achievement levels (*Basic, Proficient,* and *Advanced*) are defined only for main-NAEP reporting, whereas LTT results are still reported in terms of performance levels defined by the earlier scale anchoring procedure.

Content differences between main NAEP and the LTT stand as one of the main rationales for maintaining the LTT, and arguments have been presented for maintaining age-based sampling for the LTT as well. The case for maintaining data collections at different times of the year, however, appears much weaker. Regarding the main-NAEP digital platform transition, a strong argument can be made for also shifting the LTT to a digital platform as soon as possible.

Changing the LTT testing windows to bring them into alignment with the main-NAEP testing window would bring several advantages. First, the sampling designs for main NAEP and the LTT could then be integrated, resulting in significant cost savings. Second, these savings might make it possible at the same time to expand LTT samples somewhat to include oversampling of minority groups, thereby improving LTT estimates of achievement gaps and subgroup trends, as recommended by Barron and Koretz (1996).

Third, for technical reasons, integrating the two samples would increase the precision of contrasts between main-NAEP versus LTT gaps and trends.[37] This change might also reduce costs by enabling better synchronization of schedules for main-NAEP versus LTT data analysis and reporting.

Facility with the format and conventions of paper-and-pencil testing per se seems peripheral to the kinds of inferences that educators, policymakers, and the public at large are most likely to wish to draw from LTT performance and performance trends. Indeed, paper-and-pencil testing might become a liability. As children become increasingly accustomed to digital devices, their performance on paper-and-pencil assessments might decline for reasons unrelated to their proficiency with the content and skills assessed, including diminished motivation to write out responses long-hand and declining familiarity with hard-copy answer booklets. The future is difficult to predict, but given that the next scheduled LTT assessment is not until 2024, continued reliance on already outmoded assessment technology is a serious concern. Another argument against continued paper-and-pencil testing is the greater cost of producing, shipping, collecting, scanning, and scoring hard-copy testing materials. In this regard, NCES provided the following statement in the course of its review of this white paper:

> **Feasibility of maintaining LTT NAEP as a paper-based assessment**
>
> By the next LTT assessment administration in 2024, Main NAEP will have transitioned completely to a digitally based assessment (DBA). LTT, however, continues to be a paper based assessment (PBA). Delivering assessments via both formats would require supporting two test delivery systems with different administrative procedures. Specifically, under that scenario, various processes and procedures would need to be reinstituted solely for LTT, including booklet printing, booklet shipping, booklet processing, data transfer, and constructed-response scoring. Maintaining paper for the LTT would therefore, require an additional parallel set of resources to accomplish this work and would be cost prohibitive. Moreover, administration procedures appropriate for the separate PBA would need to be reinstituted. For example, while separate trainings would be required for Main NAEP and LTT regardless of the format of LTT, maintaining LTT as a paper-based assessment would now require two different sets of procedures for field staff training, for which separate quality control procedures would have to be developed and implemented. Given these challenges, NCES posits that maintaining a paper-based LTT in the context of a fully digital Main NAEP is largely infeasible.

These considerations give rise to recommendations for two major updates to the LTT design. One substantial, but manageable, design change would be to shift the LTT testing windows so that all LTT and main-NAEP testing occurred at the same time. A second substantial, but manageable, design change would move LTT assessments from paper-and-pencil booklets to digital assessments. These two updates might be considered independently, but in the interest of minimizing the number of bridge studies and trend line interruptions required, it seems most sensible to investigate both at the same time.

Testing windows. Under this LTT redesign proposal, sampling frames for main NAEP and the LTT would be integrated. When main-NAEP student samples were drawn, LTT student samples would also be

[37] Positive sampling error covariances between the main-NAEP and LTT components of the NAEP assessment program, arising from use of the same primary sampling units (PSUs), would increase the precision of contrasts between main-NAEP and LTT estimates.

drawn, in a subset of the same schools.[38] Slightly larger numbers of 9-year-old 4th graders, 13-year-old 8th graders, and 17-year-old 12th graders would be tested, together with supplemental samples of 9-, 13-, and 17-year-olds not at these modal grade levels. Thus, a single sampling plan would include both grade-based main-NAEP samples and age-based LTT samples. Sample integration might also enable better alignment between main-NAEP and LTT contextual variables used for conditioning estimates in the MML/multiple imputation analysis plan. As noted, particular care would be required in updating any contextual variables used in defining separately reported subgroups.

Changing the LTT testing windows would also require redefining the ranges of birthdates framing each age cohort. (For example 13-year-olds in the fall are not the same group as 13-year-olds in January through March.) Definitions of LTT age cohorts have shifted from time to time, with current definitions from January 1 to December 31 for 9-year-olds and 13-year-olds, and October 1 through September 30 for 17-year-olds. Because the current LTT testing window for 9-year-olds (January through March) matches the main-NAEP testing window, presumably, the age cohort definition for 9-year-olds would remain unchanged. Cohort definition for 13-year-olds might also remain unchanged, but the October 1 through September 30 cohort definition for 17-year-olds would require attention.[39] The way the 17-year-old age cohort is currently defined, this group's modal grade is 11, not 12. For the 2012 LTT, for example, age 9 students were defined as those born during 2002 (and were therefore 9 years old as of January 1, 2012). Age 13 students were those born during 1998. However, age 17 students were those born between October 1, 1994, and September 30, 1995, a shift of nine months relative to the calendar-year definition that would correspond most closely to the definitions used for ages 9 and 13. Given that LTT assessments occur in the fall for 13-year-olds and the following spring for 17-year-olds, there is a justification for this disparity. However, with the shift to a common testing window, these discrepant age cohort definitions might be changed.

Transition to digital platform. Under this LTT redesign proposal, LTT exercises would be migrated to a digital platform. This change would be planned so as to minimize changes to exercises. For most modern assessments, digital testing may be viewed as an affordance for greater flexibility in assessing a broader range of cognitive skills using new, often interactive, item formats (Bennett, 2015; Drasgow, Luecht, and Bennett, 2006; Expert Panel on the Future of NAEP, 2012). For the LTT, however, the digital platform would in effect be little more than an "electronic page-turner," that is, a device for presenting stimuli and recording responses using items originally designed for paper-and-pencil tests. In addition to new item formats, digital platforms also support significant improvements to testing accommodations for students with disabilities and possibly for English-language learners, better approaching the ideal of "universal design" (Way, Davis, Keng, and Strain-Seymour, 2016). Even at the risk of diminished comparability with past assessment results, making such improvements to available accommodations as part of a digitally based LTT would seem to be a sound investment in more accessible and thereby more valid measurement for the broadest possible range of students, as well as enhanced comparability with main-NAEP assessment results. Issues to consider in the transition of the LTT to a digital platform would

[38] LTT samples need only be large enough to estimate achievement distributions for populations and subgroups at the national level, whereas main-NAEP, at least at grades 4 and 8, also provides estimates at the state level. It is possible that LTT sampling might require some slight augmentation of the pool of main-NAEP schools to include students within LTT age definitions but outside modal grade levels. Because state-level reporting for grades 4 and 8 already requires larger NAEP samples, such augmentation to the pool of main-NAEP schools is most likely to be required for 17-year-olds.

[39] Explicit LTT age cohort definitions may be found at https://nces.ed.gov/nationsreportcard/ltt/sampledesign.aspx.

include typing (versus writing) extended responses, especially for 9-year-olds, and mathematical computations possibly using scratch paper.

A bridge study would be required, involving the testing of at least two groups of 9-year-olds, at least two groups of 13-year-olds, and at least two groups of 17-year-olds. At ages 13 and 17, some students would be tested during the current LTT testing window and the remainder during the January-March main-NAEP testing window. For LTT 9-year-olds and main-NAEP fourth graders, the testing windows already coincide, and so the two LTT bridge study 9-year-old samples would be randomly equivalent. At age 13, the samples would be drawn using identical (calendar-year) age cohort definitions, but might differ subtly due to patterns of student mobility over the course of the school year. Note that mobility almost certainly affects demographic subgroups differentially. Mobility effects, as well as the effects of maturation and schooling between the fall LTT testing window and the January-March main-NAEP testing window, would be absorbed into the mix of effects accounted for by the bridge study. At age 17, there would be a nine-month offset in the range of birthdates included in the respective sampling frames. This effect, along with effects of mobility and high school dropout, as well as procedural changes, would all be confounded in the effects estimated by the bridge study. Acknowledging the impact on LTT trend lines, the bridge study would support a transition to efficient, workable LTT sampling frames for 13- and 17-year-olds, as well as 9-year-olds, all tested during the main-NAEP testing window.

In addition, the bridge study would test some students at each age level using paper-and-pencil booklets and others using the digital platform. The simplest bridge study would use just two groups. One group would be tested using paper-and-pencil booklets during the current LTT testing window and the other would be tested using a digital platform during the main-NAEP testing window. This design would permit estimation of the combined effect of changes to both the testing window (current LTT versus main-NAEP) and platform (paper-and-pencil versus digital). A more sophisticated design would employ four groups at ages 13 and 17, tested using each possible combination of testing window and platform. This design would enable separate estimation of the effects of each change, as well as potential interactions between them. A third bridge study design might be a compromise, using just two groups at age 9, tested with paper-and-pencil booklets (i.e., continuing the LTT status quo) and using a digital platform; and three groups each at ages 13 and 17, tested during the current LTT testing window with paper-and-pencil booklets, during the current main-NAEP testing window with paper-and-pencil booklets, and during the current main-NAEP testing window using a digital platform. This three-group design would provide significantly more information than the two-group design while avoiding the need to support schools in fielding digital assessments at different times of the year.

Bridge study timing. The next LTT assessment is now scheduled for fall 2023 through spring 2024. The scheduled main-NAEP assessments closest to this LTT assessment date would be conducted in January-March 2023 and January-March 2025. Historically, bridge studies have coincided with a regularly scheduled assessment data collection. Because these bridge studies have always focused on procedural changes within either main NAEP or the LTT, this complexity in scheduling a bridge study has not arisen before.[40] As classroom use of digital media increases, and as digital testing is more widely adopted for assessment purposes, children's familiarity with the routines of standardized paper-and-pencil testing may decline rapidly. Students who will be tested as 12th graders in 2023 or as 17-year-olds in 2024 are now close to the very beginning of their formal schooling. By 2024, paper-and-pencil standardized tests may

[40] Thanks to Governing Board staff for alerting the author to this concern.

be essentially obsolete. Thus, there is some urgency in scheduling the proposed LTT bridge study well before the next regular LTT assessment is scheduled to occur.

When a bridge study is conducted in conjunction with a regular NAEP data collection, most of the students included in the bridge study, those in the "new procedures" group, are also being tested as part of that regular data collection. Thus, the additional data collection cost is for a (typically) smaller sample of students tested following old procedures. Because the "new procedures" group is typically much larger, the size of "old procedures" groups is a limiting factor on the precision of bridge study findings. It follows that scheduling a bridge study independent of a regular data collection would have implications for both cost and precision. Cost and administrative burden are of course significant considerations, but a stand-alone LTT bridge study conducted as soon as possible, presumably within the next five years or so, might be seriously considered. Findings from such a study would be noisy (due to limited sample size) but might nonetheless be of considerable value in providing some information on LTT trends during the 12-year interval between the most recent LTT assessment in 2011-12 and the next scheduled LTT assessment in 2023-24. (The bridge study might be designed to estimate procedural effects for reporting subgroups as well as the full population, but trend line data points might be reported only for the nation as a whole.)

Redesigned LTT Model

The bridge study just described would enable continued use of existing LTT exercises, already organized into blocks and combined into booklets, as was done for the 2004, 2008, and 2012 LTT assessments, but migrated onto the digital platform and administered during the main-NAEP testing window. Going forward, any minor updates required would be managed in the same way as has been done for the three most recent LTT assessments. Note that during the integrated main-NAEP plus LTT data collection, each student would respond to just one booklet, providing data on either main NAEP or LTT but never both. Thus, this design could be described as employing two separate (although integrated) student samples for main NAEP versus LTT.[41] The goal would be to enable main-NAEP and LTT data collections to occur simultaneously, with students sitting in the same room at the same time, some responding to main-NAEP and others to LTT exercises. The differences would be invisible to respondents.

An option considered and rejected. Once it is envisioned that main-NAEP and LTT data collections might occur at the same time, the possibility might be considered that, rather than having some students respond solely to LTT exercises and others solely to main-NAEP exercises, LTT and main-NAEP exercise blocks might be paired in the same booklets. Such a further integration of main-NAEP and LTT would permit estimation of individual-level correlations between the LTT and main-NAEP score scales. This option was rejected, however, for the following reasons. First, LTT booklets are still configured with three 15-minute exercise blocks, whereas main-NAEP booklets feature two 25-minute blocks. Thus, reconfiguration of LTT blocks would need to be incorporated as another procedural change in the proposed near-term LTT bridge study.[42] Second, even if LTT blocks were reconfigured to the 25-minute format, the "look and feel" of LTT blocks might be quite different from main-NAEP blocks, if only due to use of more innovative exercise formats and better graphics as main-NAEP more fully exploits the affordances of the digital platform. These differences could potentially induce context effects across blocks within booklets, threatening main-NAEP trend lines. Distortions of LTT response data might be

[41] If a stand-alone bridge study were conducted before 2024, this fully integrated sampling and test administration would not occur until later.

[42] For reading, reconfiguration of 15-minute blocks to 25-minute blocks might be difficult due to shorter LTT reading passage lengths and the need to keep all exercises associated with a given passage together.

even worse, as students who have just completed an interesting, colorfully illustrated, and interactive main-NAEP exercise block might show diminished motivation when the next exercise block they encountered called for multiple-choice responses to exercises presented as text with, at most, relatively crude, static line drawings by way of illustration. Third, main-NAEP and LTT booklets include somewhat different contextual (background) questions, which are used to create conditioning variables in the course of estimating NAEP scale score distributions. While there may be reasons to update LTT contextual variables (ideally maintaining unchanged those that define reporting groups), yoking together the main-NAEP and LTT contextual questions would introduce an avoidable constraint. Fourth, this option would raise the problem of how to configure LTT booklets for students not at the modal grade level and therefore not part of the main-NAEP sample. Using different stimulus materials for off-modal-grade versus at-modal-grade LTT respondents could differentially affect the responses of these two subgroups of LTT students, distorting trends and complicating the use of LTT data for any research purposes. The alternative of including main-NAEP blocks in LTT booklets for off-modal-grade students but then not analyzing or using students' responses to the main-NAEP exercises would not only be an expensive waste of time, but might also raise flags in the Office of Management and Budget (OMB) clearance process for the NAEP data collection. These challenges are not insurmountable, but taken together, they militate against an individual-level linkage between main NAEP and the LTT as part of the regular, recurrent NAEP data collections, even after eliminating the barriers of separate testing windows and distinct testing modes (paper-and-pencil versus digital). Such linkage might instead be the subject of a special study at some point.

Combining LTT, Main NAEP, and Bridge Study Data to Support Longitudinal Research

As discussed throughout this paper, various procedural changes in NAEP have been, and will continue to be, unavoidable. Such changes have included updates to content frameworks, exercises, sampling frames, conditioning variables, decision rules and procedures for arriving at racial/ethnic and other demographic classifications, permissible accommodations, and procedures followed in data collection, analysis, and reporting. The potential effects of significant procedural changes are investigated via bridge studies. Whenever possible, small uniform effects are absorbed into scaling constants so that reporting scales are unchanged and trend lines are continuous. When the effects of some change cannot be simply adjusted for, then two sets of results are reported for the year in which changes are adopted.

At the same time, the NAEP program has a long history of efforts to promote greater access and wider use of assessment data by making statistical tools and other kinds of supports available. The ETS proposal in the early 1980s promised that among other improvements, the new design would make available "public USEFUL tapes," not just "public use tapes" (Messick, Beaton, and Lord, 1983, p. 34, capitalization in original). Efforts to make good on that promise included the *NAEP Primer* (Beaton and Gonzalez, 1995) featuring a self-weighting subsample of NAEP data, and an updated *Primer* with data from the 2005 mathematics assessment.[43] More recently, NAEP data have been made more available and secondary analysis has been simplified via the powerful and user-friendly NAEP Data Tool, and most recently, the NAEP Data Explorer accessible via the NCES website.[44]

[43] Available at https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011463.
[44] Available at http://nces.ed.gov/nationsreportcard/naepdata.

Following in that same tradition, NCES, probably via NAEP contractors, might exploit bridge study data, main-NAEP and LTT assessment data, and possibly data from one or more small, additional special data collections, as needed, to model hypothetical data collections employing current assessment procedures but conducted at earlier points in time. Each such model would yield a simulated set of individual-level plausible values, which could then be used to project current trend lines backward in time. To the extent possible, such models would incorporate all of the same background variables used for conditioning in the most recent NAEP analyses, enabling flexible use of the simulated plausible value data sets to study and compare the projected performance of various subgroups with minimal bias. The result would be a resource of great value to policy researchers and other social scientists using NAEP trend data to study achievement over time. Such a modeling effort would explicitly acknowledge that the constructs measured by NAEP assessments have evolved, consistent with the vision articulated by the Governing Board (1996; 2002) and also by the Expert Panel on the Future of NAEP (2012). Despite this evolution, the precise constructs defined by successive NAEP frameworks are extremely highly correlated, and so it is meaningful and appropriate to use measurements of one of these distinct constructs as a predictor of another.

As an example, consider the projection of the LTT age 9 reading trend line backward from 2004 to 1999. Currently, reading in 1999 and 2004 can be directly compared, and reading in 2004 and 2008 can be directly compared, but the 1999-2004 comparison is for the LTT under old procedures, including population definitions that implicitly incorporate old accommodation policies, and the 2004-08 comparison is for the LTT under new procedures, including greater inclusion of students with disabilities under more recent accommodation policies. Findings from 1999 and 2008 cannot be compared directly, largely due to these population changes. A backward projection from 2004 to 1999 would yield simulated 1999 data for the more inclusive population, which could be directly compared with LTT assessment results from 2004 and thereafter. Data from the 2004 bridge study could be used to model the effects of the 2004 procedural changes as a function of individual student characteristics, and this model could then be applied to the 1999 data to estimate a new set of plausible values for each 1999 respondent. At the same time, the 1999 sample could be augmented with additional, hypothetical respondents standing in for students excluded in 1999 due to disabilities that would not have disqualified them from participating under the newer testing accommodation policies. Finally, plausible values would be drawn from each student's modeled (hypothetical) posterior distribution. These hypothetical plausible values would have somewhat larger variances than the plausible values actually generated in 1999, reflecting uncertainty due to modeling. These newly estimated, hypothetical 1999 plausible values could then be used to estimate performance distributions on the age 9 LTT reading scale under policies and procedures introduced in 2004, for the more inclusive population as a whole as well as for traditional reporting subgroups.

If such a methodology could be perfected, it might in principle be used to extend current main-NAEP trend lines backward in time to the earliest days of NAEP. The uncertainty in such estimates would be considerable, but that uncertainty could be quantified, and the resulting continuous data series could be of considerable value for research and policy. Of course, some cautions are in order. In addition to quantifying uncertainty, it would be critically important to document, and to the extent possible, to quantify, potential sources of bias. Different interpretations of such extrapolations would entail different assumptions, limiting the kinds of interpretations that could be supported. In particular, some intended inferences might rely on the clearly untenable assumption that extrapolated results quantified historical mastery of current main-NAEP content that was in fact rarely taught in earlier years. It must be emphasized that these backward projections would not replace current main-NAEP and LTT reporting.

They would probably be produced after some significant time delay, and would be intended for research purposes.

Summary and Conclusion

The LTT is a critical component of NAEP, adding substantial value not only historically but right up to the present. As attested by current controversies surrounding state adoptions of the Common Core State Standards, there is an abiding interest on the part of some policymakers, educators, and segments of the public in students' performance on content and skills viewed as simpler and more traditional than those espoused in current curriculum reforms. This is the kind of material tested by the LTT. In addition, as the mean age of children at each grade level gradually increases over time, age-based trends offer important context for interpreting grade-based trends; and some important policy questions have been cited that can only be investigated using age-based, as opposed to grade-based, trend information. Some differences between LTT and main-NAEP data collections have persisted as a consequence of design choices made almost a half-century ago and maintained since then to avoid disrupting trends. Some of these choices, such as paced-tape booklets and "I don't know" answer choices, were eliminated in the 2004 LTT bridge study, but separate testing windows have been maintained.

Careful consideration of the uses and interpretations of the LTT suggests that the most important distinctions between the LTT and main NAEP center on the distinct content assessed and on age-based versus grade-based sampling. Differences in testing windows appear unimportant. In addition, large-scale paper-and-pencil testing is already becoming antiquated, and so LTT migration to a digital platform is a matter of some urgency. Thus, this paper proposes consideration of another LTT bridge study to continue LTT trends with digital testing, concurrent with main-NAEP data collections. This would enable cost savings via integration of main-NAEP and LTT sampling and data collection with essentially no risk to main-NAEP trend lines, preserving the legacy of trend lines reaching back to NAEP's earliest days and supporting continuation of the LTT into the future, as mandated by current legislation. When main NAEP is fully digital, the cost of a separate paper-based LTT data collection will probably be prohibitive. For reasons already discussed, this proposed LTT bridge study should be scheduled as soon as possible.

One final proposal included in this paper sketches an approach to enhancing the utility of both main-NAEP and LTT data for research purposes by projecting recent trend lines backward in time. To accomplish this, bridge study data would be used to model the effects of procedural changes at the individual student level, and the resulting models would then be applied to assessment data from prior years to generate new sets of plausible values, simulating the responses of these historical respondents under future testing conditions. The resulting data sets would be for research only, and would have to be treated with considerable caution. They would certainly not supplant current procedures for NAEP reporting.

References

Allen, N. L., McClellan, C. A., and Stoeckel, J. J. (2005). *NAEP 1999 Long-Term Trend technical analysis report: Three decades of student performance* (NCES 2005–484). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. (Available from ERIC Document Reproduction Center, No. ED485199)

Barron, S. I., and Koretz, D. M. (1996). An evaluation of the robustness of the National Assessment of Educational Progress trend estimates for racial ethnic subgroups. *Educational Assessment*, *3*(3), 209-248.

Beaton, A. E. (1987). Overview of part II: the NAEP 1983-84 data analysis. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (Report No. 15-TR-20, pp. 225-238). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED288887)

Beaton, A. E. (1990a). Introduction. In A. E. Beaton and R. Zwick (Eds.), *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21, pp. pp. 1-13). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED322216)

Beaton, A. E. (1990b). Introduction. In E. G. Johnson, and R. Zwick (Eds.), *Focusing the new design: the NAEP 1988 technical report* (Report No. 19-TR-20, pp. 3-9). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED325496)

Beaton, A. E., and Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191-204.

Beaton, A. E., and Chromy, J. R. (2010, December). *NAEP trends: Main NAEP vs. Long-Term Trend* (Paper commissioned by the NAEP Validity Studies [NVS] Panel). San Mateo, CA: NAEP Validity Studies (NVS), American Institutes for Research. (Available from ERIC Document Reproduction Center, No. ED514137)

Beaton, A. E., and Gonzalez, E. J. (1995). *The NAEP primer*. Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy. (Available from ERIC Document Reproduction Center, No. ED404374)

Beaton, A. E., and Johnson, E. G. (2004). Emerging technical innovations in NAEP. In L. V. Jones and I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 449-466). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Beaton, A. E., and Zwick, R. (1990). *The effect of changes in the National Assessment: disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress. (Available from ERIC Document Reproduction Center, No. ED322216)

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*, 370-407.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay Company.

Bourque, M. L., and Byrd, S. (Eds.) (2000). *Student performance standards on the National Assessment of Educational Progress: Affirmations and improvements*. Washington, DC: National Assessment Governing Board. (Downloaded from https://www.nagb.org/content/nagb/assets/documents/publications/achievement/naep-student-performance-standards-affirmation-improvements.pdf)

Braun, H. I, and Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, *86*(7), 488-497.

Campbell, J. R., Lazer, S. and Mullis, I. V. S. (1994). Developing the NAEP objectives, items, and background questions for the 1992 assessments of reading, mathematics, and writing. In E. G. Johnson and J. E. Carlson (Eds.), *The NAEP 1992 technical report* (Report No. 23-TR20, pp. 33-66). Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education. (Available from ERIC Document Reproduction Center, No. ED376191)

Chang, H., Donoghue, J. R., Worthington, L. H., Wang, M., and Freund, D. S. (1996). Data analysis for the long-term trend reading assessment. In N. L. Allen, D. L. Kline, and C. A. Zelenak (Eds.) (1996). *The NAEP 1994 technical report* (Report No. NCES 97-897, pp. 357-371). Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education. (Available from ERIC Document Reproduction Center, No. ED404377)

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*(8), 672-683.

Cronbach, L. J. (2004). An interview with Lee J. Cronbach. In L. V. Jones and I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 139-153). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Deming, D., and Dynarski, S. (2008). The lengthening of childhood. *Journal of Economic Perspectives*, *22*(3), 71-92.

Dickinson, E., Taylor, L., Koger, M., Moody, R., Deatz, R., and Koger, L. (2006). *Alignment of Long Term Trend and main NAEP* (Report No. FR-06-24, MOBIS Contract No. GS-10F-0087J). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. L., and Chambers, D. L. (1988). *The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 national assessment* (Report No. 17-M-01). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED300206)

Drasgow, F., Luecht, R. M., and Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: American Council on Education/Praeger.

Expert Panel on the Future of NAEP. (2012, May). *NAEP: Looking ahead, leading assessment into the future* (Recommendations to the Commissioner, National Center for Education Statistics). (Available at http://nces.ed.gov/nationsreportcard/pdf/future_of_naep_panel_white_paper.pdf)

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519-521.

Glaser, R., and Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in system development* (pp. 419-474). New York: Holt, Rinehart, and Winston.

Haertel, E. H., and Mullis, I. V. S. (1996). The evolution of the National Assessment of Educational Progress: Coherence with best practice. In J. B. Baron and D. P. Wolf (Eds.), *Performance-Based Student Assessment: Challenges and Possibilities* (Ninety-Fifth Yearbook of the National Society for the Study of Education, Part 1, pp. 287-304). Chicago: National Society for the Study of Education (distributed by the University of Chicago Press).

Johnson, E. G. (1988). Mathematics data analysis. In A. E. Beaton (Ed.), *Expanding the new design: the NAEP 1985-86 technical report* (pp. 215-240, Report No. 17-TR-20). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED355248)

Jones, L. V. (1996). A History of the National Assessment of Educational Progress and Some Questions About Its Future. *Educational Researcher*, *25*(7), 15-22.

Jones, L. V. (2004). Chronology, 1963-2003. In L. V. Jones and I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 11-21). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Jones, L. V., and Olkin, I. (Eds.) (2004). *The nation's report card: evolution and perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kolen, M. J. (2000). Issues in phasing out trend NAEP. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, and L. R. Jones (Eds.), *Grading the nation's report card: research from the evaluation of NAEP* (pp. 132-151). Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press. (Downloaded from http://www.nap.edu/catalog/9751/grading-the-nations-report-card-research-from-the-evaluation-of)

Lehmann, I. J. (2004). The genesis of NAEP. In L. V. Jones and I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 25-92). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Merwin, J. C., and Womer, F. B. (1969). Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In R. W. Tyler (Ed.), *Educational evaluation: new roles, new means* (68th Yearbook of the National Society for the Study of Education, Part II, pp. 305-334). Chicago: University of Chicago Press.

Messick, S., Beaton, A. E., and Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Rep. 83-1). Princeton, NJ: National Assessment of Educational Progress. (Available from ERIC Document Reproduction Center, No. ED236156)

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R. J. (1987). The reading data analysis: introduction. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (Report No. 15-TR-20, pp. 239-244). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED288887)

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R. J., Johnson, E. G., and Muraki, E. (1992). Scaling procedures in the National Assessment for Educational Progress. *Journal of Educational Statistics*, *17*, 131- 154.

Mislevy, R. J., and Sheehan, K. M. (1987). Trend analysis. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (Report No. 15-TR-20, pp. 361-379). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED288887)

Mislevy, R.J., and Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57-75.

Mullis, I. V. S. (1992). Developing the NAEP Content-Area Frameworks and Innovative Assessment Methods in the 1992 Assessments of Mathematics, Reading, and Writing. *Journal of Educational Measurement*, *29*(2), 111-131.

Mullis, I. V. S., and Jenkins, L. B. (1990). *The reading report card, 1971-88: Trends from the nation's report card* (Report No. 19-R-01). Princeton, NJ: National Assessment of Educational Progress at Educational Testing Service. (Available from ERIC Document Reproduction Center, No. ED315728)

Nagaoka, J., and Roderick, M. (2004). *Ending social promotion: the effects of retention*. Chicago, IL: Consortium on Chicago School Research. (Available at https://consortium.uchicago.edu/sites/default/files/publications/p70.pdf)

National Assessment Governing Board [Governing Board]. (1996). *Redesigning the National Assessment of Educational Progress: policy statement*. Washington, DC: Author. (Downloaded from https://www.nagb.org/content/nagb/assets/documents/policies/Redesigning%20the%20National%20Assessment%20of%20Educational%20Progress.pdf)

National Assessment Governing Board [Governing Board]. (2002). *Long-Term Trend Policy Statement*. Washington, DC: Author. (Downloaded from https://www.nagb.org/content/nagb/assets/documents/policies/Long-term%20Trend.pdf)

National Assessment Governing Board [Governing Board]. (2012). *Mathematics framework for the 2013 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, U. S. Department of Education. (Available at https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/mathematics/2013-mathematics-framework.pdf)

National Assessment Governing Board [Governing Board]. (2015). *Reading framework for the 2015 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, U. S. Department of Education. (Available at https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading/2015-reading-framework.pdf)

National Assessment of Educational Progress [NAEP]. (1981a). *Mathematics objectives: 1981-82 assessment* (NAEP Report No. 13-MA-10). Denver, CO: Education Commission of the States. (Available from ERIC Document Reproduction Service, ERIC No. ED211352)

National Assessment of Educational Progress [NAEP]. (1981b). *Procedural handbook: 1979-80 reading and literature assessment* (NAEP Report No. 11-RL-40). Denver, CO: Education Commission of the States. (Available from ERIC Document Reproduction Service, ERIC No. ED210300)

National Assessment of Educational Progress [NAEP]. (1985). *The Reading Report Card: Progress Toward Excellence in Our Schools. Trends in Reading over Four National Assessments, 1971-1984* (Report No. 15-R-01). Princeton, NJ: Educational Testing Service. (Available from ERIC Document Reproduction Service, ERIC No. ED264550)

National Center for Education Statistics (2013). *The Nation s Report Card: Trends in Academic Progress 2012* (Report No. NCES 2013-456). Washington, DC: Institute of Education Sciences, U.S. Department of Education. (Downloaded from http://nces.ed.gov/nationsreportcard/subject/publications/main2012/pdf/2013456.pdf)

National Council of Teachers of Mathematics [NCTM]. (2004, January/February). Rise in NAEP math scores coincides with NCTM standards. *News Bulletin*. (Available at https://web.archive.org/web/20071112221438/http://www.nctm.org/news/release.aspx?id=766)

Olson, J. F., and Goldstein, A. A. (1997). *The Inclusion of Students With Disabilities and Limited English Proficient Students in Large Scale Assessments: A Summary of Recent Progress* (Report No. NCES 97-482). Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement. (Available at https://nces.ed.gov/pubs97/97482.pdf)

Pellegrino, J. W., Jones, L. R., and Mitchell, K. J. (Eds.) (1999). *Grading the nation's report card: evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press. (Available at https://www.nap.edu/catalog/6296/grading-the-nations-report-card-evaluating-naep-and-transforming-the)

Perie, M., Moran, R., and Lutkus, A. D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005-464). U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office. (Downloaded from http://nces.ed.gov/nationsreportcard/pdf/main2005/2005464.pdf)

Ravitch, D. (2009). *To be a member of the Governing Board* (Paper commissioned for the 20th anniversary of the National Assessment Governing Board 1988-2008). Washington, DC: National Assessment Governing Board. (Available at https://www.nagb.org/content/nagb/assets/documents/who-we-are/20-anniversary/ravitch-formatted.pdf)

Stedman, L. C. (2009). *The NAEP Long-Term Trend Assessment: A Review of Its Transformation, Use, and Findings* (Paper Commissioned for the 20th Anniversary of the National Assessment Governing Board 1988–2008). Washington, DC: National Assessment Governing Board. (Available from the ERIC Document Reproduction Service, ERIC No. ED509383)

U. S. Department of Education. (2005, July 14). Spellings hails new national report card results: Today's news "proof that No Child Left Behind is working." Press release. (Available at https://web.archive.org/web/20080611090351/http://www.ed.gov/news/pressreleases/2005/07/07142005.html)

Vinovskis, M. A. (1998). *Overseeing the nation's report card: the creation and evolution of the National Assessment Governing Board* (Paper prepared for the National Assessment Governing Board [Governing Board]). Washington, DC: National Assessment Governing Board, U. S. Department of Education. (Downloaded from https://www.nagb.org/content/nagb/assets/documents/publications/95222.pdf)

Way, W. D., Davis, L. L., Keng, L., and Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and testing: improving educational and psychological measurement* (pp. 260-284). New York: Routledge.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, D.C.: Council of Chief State Schools Officers. (Available from the ERIC Document Reproduction Service, ERIC No. ED414305)

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph 18). Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (Available from the ERIC Document Reproduction Service No. ED440852)

Zwick, R. (Ed.) (1992a). Special Issue: National Assessment of Educational Progress. (whole issue of the *Journal of Educational Statistics*, *17*(2), 93-232)

Zwick, R. (1992b). Statistical and Psychometric Issues in the Measurement of Educational Achievement Trends: Examples From the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*(2), 205-218.

Appendix A
LTT Content Descriptions from *2011-2012 Long-Term Sample Exercise Booklet*[45] pages 8-10

Description of Long-Term Trend Mathematics

Ages 9, 13, and 17

The long-term trend mathematics assessment covers the following content topics: numbers and numeration; measurement; shape, size, and position; probability and statistics; and variables and relationships. Each test booklet consists of three content blocks of 15 minutes each.

- *Numbers and Numeration*: These exercises deal with the ways numbers are used, processed, or written. Knowledge and understanding of numeration and number concepts are assessed for whole numbers, common fractions, decimal fractions, integers, and percents. Considerable emphasis is placed on operations. Number properties and order relations are also included.

- *Measurement*: These exercises cover appropriate units; equivalence relations; instrument reading; length, weight, capacity, time, temperature, perimeter, area, and volume; nonstandard units; and precision and interpolation. A substantial number of the measurement exercises require the use and understanding of metric units.

- *Shape, Size, and Position*: These exercises measure objectives related to school geometry and concern plane and solid shapes, congruence, similarity, properties of triangles, properties of quadrilaterals, constructions, sections of solids, basic theorems and relationships, and rotations and symmetry.

- *Probability and Statistics*: These exercises assess collecting data; organizing data with tables, charts, and graphs; interpreting and analyzing data; drawing inferences; making generalizations; using basic statistics; predicting outcomes and determining combinations.

- *Variables and Relationships*: These exercises deal with the recognition of facts, definitions, and symbols of algebra; the solution of equations and inequalities; the use of variables to represent problem situations and elements of a number system; the evaluation and interpretation of functions and formulas; the graphing of points and lines in a coordinate system; and the use of exponential and trigonometric functions, and logic. Most of these exercises are at the 17-year-old level, at which students have had the opportunity to study algebra.

[45] See http://nces.ed.gov/nationsreportcard/subject/_commonobjects/pdf/demo_booklet/2011_12_sqb_ltt.pdf. "Demonstration Booklets" from the 2003-2004 and 2007-2008 LTT assessments contain nearly identical text, although the values in the "Target Percentages by Age Level" for the mathematics LTT differ slightly between the earlier booklets and the booklet from 2011-2012, and the earlier booklets do not give any percentage breakdowns for reading (see https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/ltt_demo_bklet.pdf and https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/08-sq-ltt.pdf).

For the three age levels assessed—9, 13, and 17—the percentage of test questions from each content topic is distributed as follows:

Target Percentages by Age Level

	Age 9	Age 13	Age 17
Numbers and numeration	50%	50%	44%
Measurement	19%	19%	12.5%
Shape, size, and position	12.5%	12.5%	12.5%
Probability and statistics	6%	6%	6%
Variables and relationships	12.5%	12.5%	25%

The long-term trend mathematics assessment includes the following process domains: mathematical knowledge, mathematical skill, mathematical understanding, and mathematical application.

- *Mathematical Knowledge*: Mathematical knowledge refers to the recall and recognition of mathematical ideas expressed in words, symbols, or figures. Mathematical knowledge relies, for the most part, on memory processes. It does not ordinarily require more complex mental processes. Exercises that assess mathematical knowledge require that a student recall or recognize one or more items of information. An example of an exercise involving recall would be one that asks for a multiplication fact, such as the product of five and two.

- *Mathematical Skill*: These exercises require the performance of specified tasks, such as making measurements, multiplying two fractions, performing mental computations, graphing a linear equation, or reading a table.

- *Mathematical Understanding*: Exercises that assess mathematical understanding require that a student provide an explanation, an illustration for one or more items of knowledge, or the transformation of knowledge. They do not require the application of that knowledge to the solution of a problem. An example of an exercise involving understanding is one that asks why a certain graph is not the graph of a function.

- *Mathematical Application*: Mathematical application and problem solving refer to the use of mathematical knowledge, skill, and understanding in solving both routine and nonroutine problems. Exercises that assess mathematical application and problem solving require a sequence of processes that relate to the formulation, solution, and interpretation of problems. The processes may include recalling and recording knowledge, selecting and carrying out algorithms, making and testing conjectures, and evaluating arguments and results. Exercises assessing mathematical application may vary from routine textbook problems to exercises dealing with mathematical arguments.

Description of Long-Term Trend Reading

Ages 9, 13, and 17

The long-term trend reading assessment contains a range of reading materials, from simple narrative passages to complex articles on specialized topics. The selections include brief stories, poems, and passages from textbooks and other age-appropriate reading material. Students' comprehension of these materials is assessed with both multiple-choice questions and constructed-response questions in which students are asked to provide a written response. In the long-term trend reading assessment, students are given selections in expository reading, narrative reading, and document reading. Each test booklet consists of three content blocks of 15 minutes each.[46]

The expository reading selections in the assessment consist of passages ranging from 250 words to 500 words at age 9 or to 800 words at age 17 and short paragraphs of 50 to 150 words at all ages. Students read a passage, then answer multiple-choice or constructed-response questions about the passage. The percentage of questions in the assessment allocated to expository reading varies, by age and by block, from 54 percent to 61 percent.

Similarly, the narrative reading selections in the assessment consist of passages ranging from 250 words to 500 words at age 9 or to 800 words at age 17 and short paragraphs of 50 to 150 words at all ages. Students read a passage, then answer multiple-choice or constructed-response questions about the passage. The narrative reading selections also include poetry passages of 50 to 150 words, followed by multiple-choice and constructed-response questions. The percentage of questions in the assessment allocated to narrative reading varies, by age and by block, from 14 percent to 23 percent.

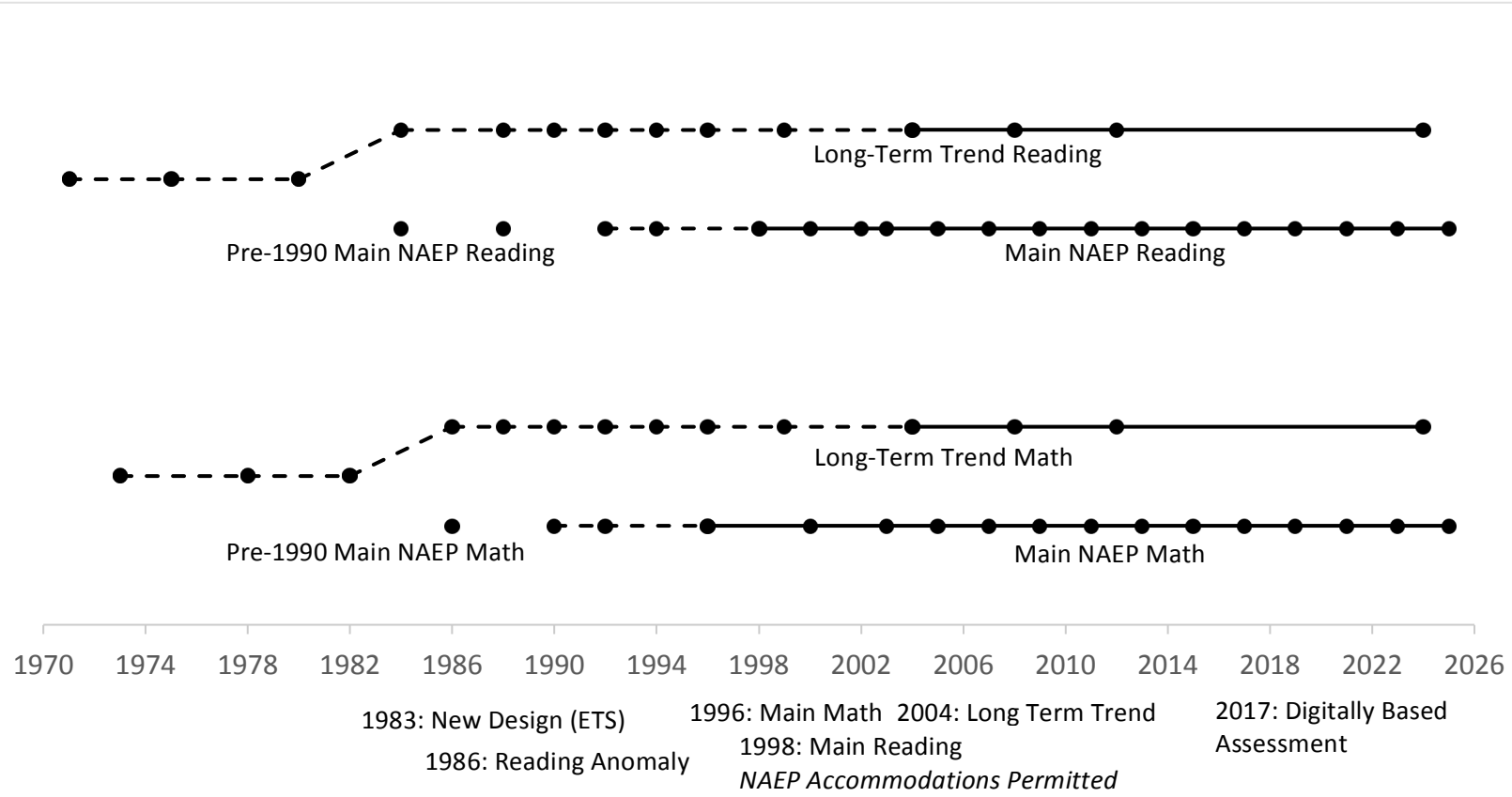
The document reading selections in the assessment consist of materials that represent real-life activities, such as a train schedule or a sale coupon. The percentage of questions in the assessment allocated to document reading varies, by age and by block, from 17 percent to 24 percent.

Percentages of Items by Text Type, Item Format, (sic) **and Age Level**

Text Type	Age 9	Age 13	Age 17
Expository	66%	59%	70%
Narrative	24%	18%	13%
Document and other	10%	23%	17%

[46] This description, taken directly from the *2011-2012 Long-Term Sample Exercise Booklet*, omits any further mention of the specific reading skills assessed by the LTT, nor does it note that most LTT reading passages were written specifically for NAEP. This contrasts with the current practice for main-NAEP of relying exclusively on authentic texts.

Figure 1: Timelines for Long-Term Trend (LTT) and Main NAEP, showing past and future anticipated trend line linkages.



Notes: Dots represent assessments. Connecting lines (either dashed or solid) indicate assessments reported on a common scale. Splits in lines in mid-1980s show beginnings of Long-Term Trends. Changes from dashed to solid lines show where trend lines were re-started (with bridge studies). Some assessments shown included only one or two age or grade levels (see https://nces.ed.gov/nationsreportcard/about/assessmentsched.asp). Schedule shown past 2017 is unofficial and subject to change.