Chapter 8
Large-Scale Group-Score Assessment

Albert E. Beaton and John L. Barone

Large-scale group assessments are widely used to inform educational policymakers about the needs and accomplishments of various populations and subpopulations. The purpose of this section is to chronicle the ETS technical contributions in this area.

Various types of data have been used to describe demographic groups, and so we must limit the coverage here. We will consider only assessments that have important measurements, such as educational achievement tests, and also have population-defining variables such as racial/ethnic, gender, and other policy-relevant variables, such as the number of hours watching TV or mathematics courses taken. The assessed population must be large, such as the United States as a whole, or an individual state.

The design of group assessments is conceptually simple: define the population and measurement instruments and then test all students in the population. For example, if a high school exit examination is administered to all high school graduates, then finding differences among racial/ethnic groupings or academic tracks is straightforward. However, if the subgroup differences are the only matter of interest, then this approach would be expensive and consume a substantial amount of student time.

To take advantage of the fact that only group and subgroup comparisons are needed, large-scale group assessments make use of sampling theory. There are two sampling areas:

- Population to be measured: Scientific samples are selected so that the population and its subpopulations can be measured to the degree required.

A.E. Beaton
Boston College, Walnut Hill, MA, USA

J.L. Barone (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: jbarone@ets.org

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_8

- Subject domain to be measured: The subject area domains may be many (e.g., reading, writing, and mathematics) and may have subareas (e.g., algebra, geometry, computational skills).

Population sampling involves selecting a sample of students that is large enough to produce estimates with sufficiently small standard errors. The domain sampling determines the breadth of measurement within a subject area. These decisions determine the costs and feasibility of the assessment.

It is informative to note the similarities and differences of group and individual assessments. Individual assessments have been in use for a long time. Some examples:

- The Army Alpha examination, which was administered to recruits in World War I.
- The *SAT*® and ACT examinations that are administered to applicants for selected colleges.

Such tests are used for important decisions about the test takers and thus must be sufficiently reliable and valid for their purposes.

As defined here, group tests are intended for population and subpopulation descriptions and not for individual decision making. As such, the tests need not measure an individual accurately as long as the target population or subpopulations parameters are well estimated.

Both group and individual assessments rely on available technology from statistics, psychometrics, and computer science. The goals of the assessment determine what technical features are used or adapted. In turn, new assessment often requires the development of enhanced technology.

For group assessments, the goal is to select the smallest sample size that will meet the assessment's measurement standards. Small subpopulations (e.g., minority students) may be oversampled to ensure a sufficient number for accurate measurement, and then sampling weights are computed so that population estimates can be computed appropriately.

Domain sampling is used to ensure that the assessment instruments cover a wide range of a subject area. Item sampling is used to create different test forms. In this way, the content of a subject-matter domain can be covered while individual students respond to a small sample of test items from the total set.

In short, group assessment typically sacrifices tight individual assessment to reduce the number of students measured and the amount of time each measured student participates in the assessment.

8.1 Organization of This Chapter

There are many different ways to present the many and varied contributions of ETS to large-scale group assessments. We have chosen to do so by topic. Topics may be considered as milestones or major events in the development of group technology. We have listed the topics chronologically to stress the symbiotic relationship of information needs and technical advancements. The information demands spur technical developments, and they in turn spur policy maker demands for information. This chapter begins by looking at the early 1960s, when the use of punch cards and IBM scoring machines limited the available technology. It leads up to the spread of large-scale group technology in use around the world.

In Sect. 8.2, Overview of Technological Contributions, 12 topics are presented. These topics cover the last half-century of development in this field, beginning with early assessments in the 1960s. ETS has had substantial influence in many but not all of these topics. All topics are included to show the contributions of other organizations to this field. Each topic is described in a few paragraphs. Some important technical contributions are mentioned but not fully described. The point here is to give an overview of large-scale group assessments and the various forces that have produced the present technology.

In Sect. 8.3, ETS and Large-Scale Assessment, gives the details of technical contributions. Each topic in Sect. 8.2 is given an individual subsection in Sect. 8.3. These subsections describe the topic in some detail. Section 8.3 is intended to be technical—but not too technical. The names of individual contributors are given along with references and URLs. Interested readers will find many opportunities to gain further knowledge of the technical contributions.

Topics will vary substantially in amount of space devoted to them depending on the degree of ETS contribution. In some cases, a topic is jointly attributable to an ETS and a non-ETS researcher.

Finally, there is an appendix, which describes in some detail the basic psychometric model used in the National Assessment of Educational Progress (NAEP). This also contains a record of the many years of comparing alternative methods for ways to improve the present methodology.

8.2 Overview of Technological Contributions

The following section is intended to give an overview of the evolving technology of large-scale group assessments. It is divided into 12 topics that describe the major factors in the development of group assessment technology. The topics are introduced chronologically, although their content may overlap considerably; for example, the topic on longitudinal studies covers 40 years. Each topic is followed by a detailed description in the next section that contains individual contributions, the

names of researchers, references, and URLs. We intend for the reader to view the Overview and then move to other sections where more detail is available.

8.2.1 Early Group Assessments

The early days of group assessments brings back memories of punch cards and IBM scoring machines. Two pioneering assessments deserve mention:

- Project TALENT: The launching of Sputnik by the Soviet Union in 1957 raised concern about the quantity and quality of science education in the United States. Were there enough students studying science to meet future needs? Were students learning the basic ideas and applications of science? To answer these and other questions, Congress passed the National Defense Education Act (NDEA) in 1958.[1] To gather more information, Project TALENT was funded, and a national sample of high school students was tested in 1960. This group assessment was conducted by the American Institutes for Research.
- IEA Mathematics Assessment: At about the same time, International Association for the Evaluation of Educational Achievement (IEA) was formed and began gathering information for comparing various participating countries.

  ETS was not involved in either of these studies.

8.2.2 NAEP's Conception

In 1963, Francis Keppel was appointed the United States Commissioner of Education. He found that the commissioner was required to report annually on the progress of education in the United States. To this end, he wrote Ralph Tyler, who was then the director of the Institute for Advanced Studies in the Behavioral Sciences, for ideas on how this might be done. Tyler responded with a memorandum that became the beginning of the NAEP.

[1] U. S. Congress. National Defense Education Act of 1958, P.L. 85-864. 85th Congress, September 2, 1958. Washington, DC: GPO.U. S. Congress. The NDEA was signed into law on September 2, 1958 and provided funding to United States education institutions at all levels.

8.2.3 Educational Opportunities Survey (EOS)

Among the many facets of the Civil Rights Act of 1964[2] was the commissioning of a survey of the equality of educational opportunity in the United States. Although the EOS study did report on various inputs to the educational system, it focused on the output of education as represented by the test scores of various racial/ethnic groups in various regions of the country. The final report of this EOS, which is commonly known as the Coleman report (Coleman et al. 1966) has been heralded as one of the most influential studies ever done in education (Gamoran and Long 2006).

ETS was the prime contractor for this study. The project demonstrated that a large-scale study could be designed, administered, analyzed, interpreted, and published in a little over a year.

8.2.4 NAEP'S Early Assessments

The first phase of NAEP began with a science assessment in 1969. This assessment had many innovative features, such as matrix sampling, administration by tape recorder, and jackknife standard error estimates. In its early days, NAEP was directed by the Education Commission of the States.

8.2.5 Longitudinal Studies

The EOS report brought about a surge of commentaries in Congress and the nation's courts, as well as in the professional journals, newspapers, and magazines (e.g., Bowles and Levin 1968; Cain and Watts 1968). Different commentators often reached different interpretations of the same data (Mosteller et al. 2010; Viadero 2006). Harvard University sponsored a semester-long faculty seminar on the equality of educational opportunity that produced a number of new analyses and commentaries (Mosteller and Moynihan 1972). It soon became apparent that more data and, in particular, student growth data were necessary to address some of the related policy questions. The result was the start of a series of longitudinal studies.

[2] Civil Rights Act of 1964, P.L. No. 88-352, 78 Stat. 241 (July 2, 1964).

8.2.6 Scholastic Aptitude Test (SAT) Score Decline

In the early 1970s, educational policymakers and the news media noticed that the average SAT scores had been declining monotonically from a high point in 1964. To address this phenomenon, the College Board formed a blue ribbon panel, which was chaired by Willard Wirtz, a former Secretary of Labor. The SAT decline data analysis for this panel required linking Project Talent and the National Longitudinal Study[3] (NLS-72) data. ETS researchers developed partitioning analysis for this study. The panel submitted a report titled *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline* (Wirtz 1977).

8.2.7 Calls for Change

The improvement of the accuracy and timeliness of large-scale group assessments brought about requests for more detailed policy information. The 1980s produced several reports that suggested further extensions of and improvement in the available data on educational issues. Some reports were particularly influential:

- The Wirtz and Lapointe (1982) report made suggestions for improvement of NAEP item development and reporting methods.
- The *Nation at Risk* report (National Commission on Excellence in Education 1983) decried the state of education in the United States and suggested changes in the governance of NAEP.

8.2.7.1 The Wall Charts

Secretary of Education, Terrence Bell, wanted information to allow comparison of educational policies in different states. In 1984, he released his *wall charts,* presenting a number of educational statistics for each state, and challenged the educational community to come up with a better state indicator of student achievement. These reports presented challenges to NAEP and other information collection systems.

[3] The National Longitudinal Study of the high school class of 1972 was the first longitudinal study funded by the United States Department of Education's National Center for Education Statistics (NCES).

8.2.8 NAEP's New Design

In 1983, the National Institute of Education released a request for proposals for the NAEP grant. ETS won this competition. The general design has been published by Messick et al. (1983) with the title, *A New Design for a New Era*. Archie Lapointe was the executive director of this effort.

Implementing the new design was challenging. The NAEP item pool had been prepared by the previous contractor, Education Commission of the States, but needed to be organized for balanced incomplete block (BIB) spiraling. Foremost was the application of item response theory (IRT), which was largely developed at ETS by Lord (see, for example, Carlson and von Davier, Chap. 5, this volume). IRT was used to summarize a host of item data into a single scale. The sample design needed to change to allow both age and grade sampling. The sample design also needed to be modified for bridge studies (studies designed to link newer forms to older forms of an assessment), which were needed to ensure maintenance of existing trends.

The implementation phase brought about opportunities for improving the assessment results. The first assessment under the new design occurred in the 1983–1984 academic year and assessed reading and writing. A vertical reading scale was developed so that students at various age and grade levels could be compared. Scale anchoring was developed to describe what students knew and could do at different points on the scale. Since the IRT methods at that time could handle only right/wrong items, the average response method (ARM) was developed for the writing items, which had graded responses. The approach to standard errors using the jackknife method used replicate weights to simplify computations using standard statistical systems.

The implementation was not without problems. It was intended to use the LOGIST program (Wood et al. 1976) to create maximum likelihood scores for individual students. However, this method was unacceptable, since it could not produce scores for students who answered all items correctly or scored below the chance level. Instead, a marginal maximum likelihood program (BILOG; Mislevy and Bock 1982) was used. This method produced a likelihood distribution for each student, and five plausible values were randomly chosen from those distributions. Mislevy (1985) has shown that plausible values can produce consistent estimates of group parameters and their standard errors.

Another problem occurred in the 1985–1986 NAEP assessment, in which reading, mathematics, and science were assessed. The results in reading were anomalous. Intensive investigations into the reading results produced a report by Beaton and Zwick (1990).

ETS's technical staff has continued to examine and improve the assessment technology. When graded responses were developed for IRT, the PARSCALE program (Muraki and Bock 1997) replaced the ARM program for scaling writing data. Of special interest is the examination of alternative methods for estimating population

distributions. A detailed description of alternative methods and their evaluation is provided in the appendix.

The introduction of IRT into NAEP was extremely important in the acceptance and use of NAEP reports. The 1983–1984 unidimensional reading scale led the way and was followed by multidimensional scales in mathematics, science, and reading itself. These easy to understand and use scales facilitated NAEP interpretation.

8.2.9 NAEP's Technical Dissemination

Under its new design, NAEP produced a series of reports to present the findings of completed assessments. These reports were intended for policymakers and the general public. The reports featured graphs and tables to show important findings for different racial/ethnic and gender groupings. The publication of these reports was announced at press conferences, along with press releases. This method ensured that NAEP results would be covered in newspapers, magazines, and television broadcasts.

NAEP has also been concerned with describing its technology to interested professionals. This effort has included many formal publications:

- *A New Design for a New Era* (Messick et al. 1983), which describes the aims and technologies that were included in the ETS proposal.
- Textual reports that described in detail the assessment process.
- Descriptions of NAEP technology in professional journals.
- Research reports and memoranda that are available to the general public.
- A NAEP Primer that is designed to help secondary analysts get started in using NAEP data.

The new design included public-use data files for secondary analysis, and such files have been prepared for each NAEP assessment since 1983. However, these files were not widely used because of the considerable intellectual commitment that was necessary to understand the NAEP design and computational procedures. To address the need of secondary analysts, ETS researchers developed a web-based analysis system, the NAEP Data Explorer, which allows the user to recreate the published tables or revise them if needed. The tables and the associated standard errors are computed using the full NAEP database and appropriate algorithms. In short, powerful analyses can be computed using simple commands.[4]

This software is necessarily limited in appropriate ways; that is, in order to protect individual privacy, the user cannot identify individual schools or students. If a table has cells representing very small samples, the program will refuse to compute the table. However, the database sample is large, and such small cells rarely occur.

For more sophisticated users, there is a series of data tools that help the user to select a sample that is appropriate for the policy question at issue. This program can

[4] This software is freely available at http://nces.ed.gov/nationsreportcard/naepdata/

produce instructions for use with available statistical systems such as SAS or SPSS. For these users, a number of programs for latent regression analyses are also provided. These programs may be used under licenses from ETS.

8.2.10 National Assessment Governing Board

The National Assessment Governing Board was authorized by an amendment to the Elementary and Secondary Education Act in 1988. The amendment authorized the Governing Board to set NAEP policies, schedules, and subject area assessment frameworks. The governing board made important changes in the NAEP design that challenged the ETS technical staff.

The major change was allowing assessment results to be reported by individual states so that the performance of students in various states could be compared. Such reporting was not permitted in previous assessments. At first, state participation was voluntary, so that a sample of students from nonparticipating states was needed to provide a full national sample. ETS ran several studies to assess the effects of changing from a single national sample to national data made up from summarizing various state results.

Comparing state results led to concern about differing states exclusion procedures. NAEP had developed tight guidelines for the exclusion of students with disabilities or limited English ability. However, differing state laws and practices resulted in differences in exclusion rates. To address this problem, two different technologies for adjusting state results were proposed and evaluated at a workshop of the National Institute of Statistical Sciences.

The No Child Left behind Act (2002) required that each state provide standards for student performance in reading and mathematics at several grade levels. Using NAEP data as a common measure, ETS studied the differences in the percentages of students at different performance levels (e.g., proficient) in different states.

On another level, the Governing Board decided to define aspirational achievement levels for student performance, thus replacing the scale anchoring already in practice in NAEP. ETS did not contribute to this project; however, the method used to define aspirational levels was originally proposed by William Angoff, an ETS researcher.

At around the same time, ETS researchers looked into the reliability of item ratings (ratings obtained through human scoring of open-ended or constructed student responses to individual assessment items).This resulted in a review of the literature and recommendations for future assessments.

ETS has also explored the use of computer-based assessment models. This work used models for item generation as well as item response evaluation. An entire writing assessment was developed and administered. The possibilities for future assessments are exciting.

The appropriateness of the IRT model became an important issue in international assessments, where different students respond in different languages. It is possible

that the IRT models will fit well in one culture but not in another. The issue was faced directly when Puerto Rican students were assessed using NAEP items that were translated into Spanish. The ETS technical staff came up with a method for testing whether or not the data in an assessment fit the IRT model. This approach has been extended for latent regression analyses.

8.2.11 NAEP's International Effects

Beginning with the 1988 International Assessment of Educational Progress (IAEP) school-based assessment, under the auspices of ETS and the United Kingdom's National Foundation for Educational Research, the ETS NAEP technologies for group assessment were readily adapted and extended into international settings. In 1994, ETS in collaboration with Statistics Canada conducted the International Adult literacy Survey (IALS), the world's first internationally comparative survey of adult skills. For the past 20 years, ETS group software has been licensed for use for the Trends in International Mathematics and Science Study (TIMSS), and for the past 15 years for the Progress in International Reading Literacy Study (PIRLS). As the consortium and technology lead for the 2013 Programme for the International Assessment of Adult Competencies (PIAAC), and the 2015 Program for International Student Assessment (PISA), ETS continues its research efforts to advance group assessment technologies—advances that include designing and developing instruments, delivery platforms, and methodology for computer-based delivery and multistage adaptive testing.

8.2.12 Other ETS Technical Contributions

ETS has a long tradition of research in the fields of statistics, psychometrics, and computer science. Much of this work is not directly associated with projects such as those mentioned above. However, much of this work involves understanding and improving the tools used in actual projects. Some examples of these technical works are described briefly here and the details and references are given in the next section of this paper.

F4STAT is a flexible and efficient statistical system that made the implementation of assessment data analysis possible. Development of the system began in 1964 and has continued over many following years.

One of the basic tools of assessment data analysis is multiple regressions. ETS has contributed to this field in a number of ways:

- Exploring methods of fitting robust regression statistics using power series.
- Exploring the accuracy of regression algorithms.
- Interpreting least squares without sampling assumptions.

ETS has also contributed to the area of latent regression analysis.

8.3 ETS and Large-Scale Assessment

8.3.1 Early Group Assessments

8.3.1.1 Project Talent

Project Talent was a very large-scale group assessment that reached for a scientific sample of 5% of the students in American high schools in 1960. In the end, Project Talent collected data on more than 440,000 students in Grades 9 through 12, attending more than 1,300 schools. The students were tested in various subject areas such as mathematics, science, and reading comprehension. The students were also administered three questionnaires that included items on family background, personal and educational experiences, aspirations for future education and vocation, and interests in various occupations and activities. The students were followed up by mail questionnaires after high school graduation. ETS was not involved in this project.[5]

8.3.1.2 First International Mathematics Study (FIMS)

At about the same time, the IEA was formed and began an assessment of mathematical competency in several nations including the United States. The IEA followed up this assessment with assessments in different subject areas at different times. Although ETS was not involved in the formative stage of international assessments it did contribute heavily to the design and implementation of the third mathematics and science study (TIMSS) in 1995.[6]

8.3.2 NAEP's Conception

The original design was created by Ralph Tyler and Princeton professor John Tukey. For more detailed information see *The Nation's Report Card: Evolutions and Perspectives* (Jones and Olkin 2004).

[5] More information is available at http://www.projecttalent.org/

[6] See http://nces.ed.gov/timss/

8.3.3 Educational Opportunities Survey

The Civil Rights Act of 1964 was a major piece of legislation that affected the American educational system. Among many other things, the act required that the U.S. Office of Education undertake a survey of the equality of educational opportunity for different racial and ethnic groups. The act seemed to require measuring the effectiveness of inputs to education such as the qualifications of teachers and the number of books in school libraries. Ultimately, it evolved into what we would consider today to be a value-added study that estimated the effect of school input variables on student performance as measured by various tests. The final report of the EOS, *The Equality of Educational Opportunity* (Coleman et al. 1966), has been hailed as one of the most influential reports in American education (Gamoran and Long 2006).

The survey was conducted under the direction of James Coleman, then a professor at Johns Hopkins University, and an advisory committee of prominent educators. NCES performed the sampling, and ETS received the contract to conduct the survey. Albert Beaton organized and directed the data analysis for ETS. John Barone had key responsibilities for data analysis systems development and application. This massive project, one of the largest of its kind, had a firm end date: July 1, 1966. Mosteller and Moynihan (1972) noted that the report used data from "some 570,000 school pupils" and "some 60,000 teachers" and gathered elaborate "information on the facilities available in some 4,000 schools."

The analysis of the EOS data involved many technical innovations and adaptations: foremost, the analysis would have been inconceivable without F4STAT.[7] The basic data for the surveyed grades (Grades 1, 3, 6, 9, and 12) and their teachers' data were placed on a total of 43 magnetic tapes and computer processing took 3 to 4 hours per analysis per grade—a formidable set of data and analyses given the computer power available at the time. With the computing capacity needed for such a project exceeding what ETS had on hand, mainframe computers in the New York area were used. Beaton (1968) provided details of the analysis.

The modularity of F4STAT was extremely important in the data analysis. Since the commercially available computers used a different operating system, a module had to be written to bridge this gap. A separate module was written to enter, score, and check the data for each grade so that the main analysis programs remained the same while the modules varied. Modules were added to the main programs to create publishable tables in readable format.

The data analysis involved fitting a regression model using the variables for students, their backgrounds, and schools that was collected in the survey. The dependent variables were test scores, such as those from a reading or mathematics test. The sampling weights were computed as the inverse of the probability of selection. Although F4STAT allowed for sampling weights, the sampling weights summed to the population size, not the sample size, which inappropriately reduced the error

[7] F4STAT is described in the next section.

estimates, and so sampling errors were not published.[8] John Tukey, a professor at Princeton University, was a consultant on this project. He discussed with Coleman and Beaton the possibility of using the jackknife method of error estimation. The jackknife method requires several passes over slightly modified data sets, which was impossible within the time and resource constraints. It was decided to produce self-weighting samples of 1,000 for each racial/ethnic grouping at each grade. Linear regression was used in further analyses.

After the EOS report was published, George Mayeske of the U.S. Office of Planning, Budgeting, and Evaluation organized further research into the equality of educational opportunity. Alexander Mood, then Assistant Commissioner of NCES, suggested using commonality analysis. Commonality analysis was first suggested in papers by Newton and Spurell (1967a, b). Beaton (1973a) generalized the algorithm and detailed its advantages and limitations**.** John Barone analyzed the EOS data using the commonality technique. This resulted in books by Mayeske et al. (1972, 1973a, b), and Mayeske and Beaton (1975).

The Mayeske analyses separated the total variance of student performance into "within-school" and "among-school" components. Regressions were run separately for within- and among-school components. This approach was a precursor to hierarchical linear modeling, which came later (Bryk and Raudenbush 1992).

Criterion scaling was also an innovation that resulted from experiences with the EOS. Large-scale analysis of variance becomes tedious when the number of levels or categories is large and the numbers of observations in the cells are irregular. Coding category membership by indicator or dummy variables may become impractically large. For example, coding all of the categorical variables for the ninth-grade students used in the Coleman report would entail 600 indicator variables; including all possible interactions would involve around 10^{75} such variables, a number larger than the number of grains of sand in the Sahara Desert.

To address this problem, Beaton (1969) developed *criterion scaling*. Let us say that there is a criterion or dependent variable that is measured on a large number of students who are grouped into a number of categories. We wish to test the hypothesis that the expected value of a criterion variable is the same for all categories. For example, let us say we have mathematics scores for students in a large number of schools and we wish to test the hypothesis that the school means are equal. We can create a criterion variable by giving each student in a school the average score of all students in that school. The regression of the individual mathematics scores on the criterion variable produced the results of a simple analysis of variance. The criterion variable can be used for many other purposes. This method and its advantages and limitations were described by Pedhazur (1997), who also included a numerical example.

[8] Later, F4STAT introduced a model that made the sum of the weights equal to the sample size.

8.3.4 NAEP's Early Assessments

The early NAEP assessments were conducted under the direction of Ralph Tyler and Princeton professor John Tukey. The Education Commission of the States was the prime administrator, with the sampling and field work done by a subcontract with the Research Triangle Institute.

The early design of NAEP had many interesting features:

- Sampling by student age, not grade. The specified ages were 9-, 13-, and 17-year-olds, as well as young adults. Out of school 17-year-olds were also sampled.
- Use of matrix sampling to permit a broad coverage of the subject area. A student was assigned a booklet that required about an hour to complete. Although all students in an assessment session were assigned the same booklet, the booklets varied from school to school.
- Administration by tape recorder. In all subject areas except reading, the questions were read to the students through a tape recording, so that the effect of reading ability on the subject areas would be minimized.
- Results were reported by individual items or by the average percentage correct over various subject matter areas.
- The jackknife method was used to estimate sampling variance in NAEP's complex sampling design.

For more extensive discussion of the design see Jones and Olkin (2004).

ETS was not involved in the design and analysis of these data sets, but did have a contract to write some assessment items. Beaton was a member of the NAEP computer advisory committee. ETS analyzed these data later as part of its trend analyses.

8.3.5 Longitudinal Studies

The EOS reported on the status of students at a particular point in time but did not address issues about future accomplishments or in-school learning. Many educational policy questions required information about growth or changes in student accomplishments. This concern led to the funding and implementation of a series of longitudinal studies.

ETS has made many important contributions to the methodology and analysis technology of longitudinal assessments. Continual adaptation occurred as the design of longitudinal studies responded to different policy interests and evolving technology. This is partially exemplified by ETS contributions addressing multistage adaptive testing (Cleary et al. 1968; Lord 1971), IRT intersample cross-walking to produce comparable scales, and criterion-referenced proficiency levels as indicators of student proficiency. Its expertise has been developed by the longitudinal study group, which was founded by Thomas Hilton**,** and later directed by Donald Rock,

and then by Judy Pollack**.** We will focus here on the national longitudinal studies sponsored by the U.S. Department of Education[9].

The first of the national studies was the National Longitudinal Study of the Class of 1972[10] (Rock et al. 1985) which was followed by a series of somewhat different studies. The first study examined high school seniors who were followed up after graduation. The subsequent studies measured high school accomplishments as well as postsecondary activities. The policy interests then shifted to the kindergarten and elementary years. The change in student populations being studied shows the changes in the policymakers' interests.

Rock (Chap. 10, this volume) presented a comprehensive 4-decade history of ETS's research contributions and role in modeling and developing psychometric procedures for measuring change in large-scale longitudinal assessments. He observed that many of these innovations in the measurement of change profited from research solutions developed by ETS for NAEP.

In addition to the national studies, ETS has been involved in other longitudinal studies of interest:

- Study of the accomplishments of U.S. Air Force members 25 years after enlistment. The study (Thorndike and Hagen 1959) was done in collaboration with the National Bureau for Economic Research. Beaton (1975) developed and applied econometric modeling methods to analyze this database.
- The Parent Child Development Center (PCDC) study[11] of children from birth through the elementary school years. This study was unique in that the children were randomly assigned *in utero* to treatment or control groups. In their final evaluation report, Bridgeman**,** Blumenthal, and Andrews (Bridgeman et al. 1981) indicated that replicable program effects were obtained.

8.3.6 SAT Score Decline

In the middle of the 1970s, educational policymakers and news media were greatly concerned with the decline in average national SAT scores. From 1964 to the mid-1970s, the average score had dropped a little every year. To study the phenomenon, the College Board appointed a blue ribbon commission led by Willard Wirtz, a former U.S. Secretary of Labor.

The question arose as to whether the SAT decline was related to lower student ability or to changes in the college-entrant population. ETS researchers proposed a

[9] National Longitudinal studies were originally sponsored by the U.S. Office of Education. That office evolved into the present Department of Education.

[10] Thomas Hilton was the principal investigator; Hack Rhett and Albert Beaton contributed to the proposal and provided team leadership in the first year.

[11] Samuel Messick and Albert Beaton served on the project's steering committee. Thomas Hilton of the ETS Developmental Research Division was the Project Director. Samuel Ball and Brent Bridgeman directed the PCDC evaluation.

design to partition the decline in average SAT scores into components relating to shifts in student performance, shifts in student populations, and their interaction. To do so required that comparable national tests be available to separate the college-bound SAT takers from the other high school students. The only available national tests at that time were the tests from Project Talent and from NLS-72 . A carefully designed study linking the tests was administered to make the test scores equivalent.

8.3.6.1 Improvisation of Linking Methods

The trouble was that the reliabilities of the tests were different. The Project Talent test had 49 items and a higher reliability than the NLS-72 20-item test. The SAT mean was substantially higher for the top 10% of the Project Talent scores than of the NLS-72 scores, as would be expected from the different reliabilities. Improving the reliability of the NLS-72 test was impossible; as Fred Lord wisely noted that, if it were possible to convert a less reliable test to a reliable one, there would be no point to making reliable tests. No equating could do so.

The study design required that the two tests have equal—but not perfect—reliability. If we could not raise the reliability of the NLS-72 test, we could lower the reliability of the Project Talent test. We did so by adding a small random normal deviate to each Project Talent score where the standard deviation of the normal deviate was calculated to give the adjusted Project Talent scores the same reliability as the NLS-72 scores. When this was done, the SAT means for the top two 10% samples were within sampling error.

8.3.6.2 Partitioning Analysis

Partitioning analysis (Beaton et al. 1977) was designed for this study. Many scientific studies explore the differences among population means. If the populations are similar, then the comparisons are straightforward. However, if they differ, the mean comparisons are problematic. Partitioning analysis separates the difference between two means into three parts: proficiency effect, population effect, and joint effect. The proficiency effect is the change in means attributable to changes in student ability, the population effect is the part attributable to population changes, and the joint effect is the part attributable to the way that the population and proficiency work together. Partitioning analysis makes it simple to compute a well-known statistic, the standardized mean, which estimates what the mean would have been if the percentages of the various subgroups had remained the same.

In the SAT study, partitioning analysis showed that most of the decline in SAT means was attributable to population shifts, not changes in performance of those at particular levels of the two tests. What had happened is that the SAT-taking population had more than doubled in size, with more students going to college; that is,

democratizing college attendance resulted in persons of lower ability entering the college-attending population.

Partitioning analysis would be applied again in future large-scale-assessment projects. For example, to explore the NAEP 1985–1986 reading anomaly (discussed later in this chapter), and also in a special study and resulting paper, *Partitioning NAEP Trend Data* (Beaton and Chromy 2007), that was commissioned by the NAEP validity studies panel. The SAT project also led to a book by Hilton on merging large databases (Hilton 1992).

8.3.7 Call for Change

The early 1980s produced three reports that influenced the NAEP design and implementation:

- The Wirtz and Lapointe (1982) report *Measuring the Quality of Education: A Report on Assessing Educational Progress* commended the high quality of the NAEP design but suggested changes in the development of test items and in the reporting of results.
- The report of the National Commission on Excellence in Education (NCEE), titled *A Nation at Risk: The Imperative for Educational Reform* (NCEE 1983), decried the state of education in the United States.
- Terrence Bell, then Secretary of Education, published wall charts, which contained a number of statistics for individual states. Included among the statistics were the average SAT and ACT scores for these states. Realizing that the SAT and ACT statistics were representative of college-bound students only, he challenged the education community to come up with better statistics of student attainment.

8.3.8 NAEP's New Design

The NAEP is the only congressionally mandated, regularly administered assessment of the performance of students in American schools. NAEP has assessed proficiency in many school subject areas (e.g., reading, mathematics, science) at different ages and grades, and at times young adults. NAEP is not a longitudinal study, since individual students are not measured as they progress in schooling; instead, NAEP assesses the proficiency of a probability sample of students at targeted school levels. Progress is measured by comparing the proficiencies of eighth-grade students to students who were eighth graders in past assessments.

In 1983, ETS competed for the NAEP grant and won. Westat was the subcontractor for sampling and field operations. The design that ETS proposed is published in

A New Design for a New Era (Messick et al. 1983).[12] The new design had many innovative features:

- *IRT scaling.* IRT scaling was introduced to NAEP as a way to summarize the data in a subject area (e.g., reading). This will be discussed below.
- *BIB spiraling.* BIB spiraling was introduced to address concerns about the dimensionality of NAEP testing data. To assess a large pool of items while keeping the testing time for an individual student to less than an hour, BIB spiraling involved dividing the item pool into individually timed (e.g., 15-minute) blocks and assigning the blocks to assessment booklets so that each item is paired with each other item in some booklet. In this way, the correlation between each pair of items is estimable. This method was suggested by Beaton and implemented by James Ferris. The idea was influenced by the work of Geoffrey Beall[13] on lattice designs (Beall and Ferris 1971) while he was at ETS.
- *Grade and age ("grage") sampling*. Previous NAEP samples were defined by age. ETS added overlapping grade samples so that results could be reported either by age or by grade.
- *"Bridge" studies.* These studies were introduced to address concerns about maintaining the already existing trend data. Bridge studies were created to link the older and newer designs. Building the bridge involved collecting randomly equivalent samples under both designs.

Implementing a new, complex design in a few months is challenging and fraught with danger but presents opportunities for creative developments. The most serious problem was the inability to produce maximum likelihood estimates of proficiency for the students who answered all their items correctly or answered below the chance level. Because reading and writing blocks were combined in some assessment booklets, many students were given only a dozen or so reading items. The result was that an unacceptable proportion of students had extreme, nonestimable, reading scores. The problem was exacerbated by the fact that the proportion of high and low scorers differed by racial/ethnic groups, which would compromise any statistical conclusions. No classical statistical methods addressed this problem adequately. The maximum likelihood program LOGIST (Wingersky et al. 1982; Wingersky 1983), could not be used.

Mislevy (1985) noted that NAEP did not need individual student scores; it needed only estimates of the distribution of student performance for different subpopulations such as gender or racial/ethnic groupings. In fact, it was not permissible or desirable to report individual scores. Combining the recent developments in

[12] Archie Lapointe was executive director. Original staff members included Samuel Messick as coordinator with the NAEP Design and Analysis Committee, Albert Beaton as director of data analysis**,** John Barone as director of data analysis systems, John Fremer as director of test development, and Jules Goodison as director of operations. Ina Mullis later moved from Education Commission of the States (the previous NAEP grantee) to ETS to become director of test development.

[13] Geoffrey Beall was an eminent retired statistician who was given working space and technical support by ETS. James Ferris did the programming for Beall's work.

marginal maximum likelihood available in the BILOG program (Mislevy and Bock 1982) and the missing data theory of Rubin (1977, 1987), he was able to propose consistent estimates of various group performances.

A result of the estimation process was the production of plausible values, which are used in the computations. Although maximum likelihood estimates could not be made for some students, estimation of the likelihood of a student receiving any particular score was possible for all. To remove bias in estimates, the distribution was "conditioned" using the many reporting and other variables that NAEP collected. A sample of five plausible values was selected at random from these distributions in making group estimates. von Davier et al. (2009) discussed plausible values and why they are useful.

The development of IRT estimation techniques led to addressing another problem. At that time, IRT allowed only right/wrong items, whereas the NAEP writing data were scored using graded responses. It was intended to present writing results one item at a time. Beaton and Johnson (1990) developed the ARM to scale the writing data. Essentially, the plausible value technology was applied to linear models.

In 1988, the National Council for Measurement in Education (NCME) gave its Award for Technical Contribution to Educational Measurement to ETS researchers Robert Mislevy, Albert Beaton, Eugene Johnson, and Kathleen Sheehan for the development of the plausible values methodology in the NAEP. The development of NAEP estimation procedures over time is detailed in the appendix.

The NAEP analysis plan included using the jackknife method for estimating standard errors, as in past NAEP assessments. However, the concept of replicate weights was introduced to simplify the computations. Essentially, the jackknife method involves pairing the primary sampling units and then systematically removing one of each pair and doubling the weight of the other. This process is done separately for each pair, resulting in half as many replicate weights as primary sampling units in the full sample. The replicate weights make it possible to compute the various population estimates using a regression program that uses sampling weights.

Another problem was reporting what students in American schools know and can do, which is the purpose of the assessment. The scaling procedures summarize the data across a subject area such as reading in general or its subscales. To describe the meaning of scales, scale anchoring was developed (Beaton and Allen 1992). In so doing, several anchor points on the scale were selected at about a standard deviation apart. At each point, items were selected that a large percentage of students at that point could correctly answer and most students at the next lower point could not. At the lowest level, items were selected only on the probability of answering the item correctly. These discriminating items were then interpreted and generalized as anchor descriptors. The scale-anchoring process and descriptors were a precursor to what would become the National Assessment Governing Board's achievement levels for NAEP.

Of special interest to NAEP was the question of dimensionality, that is, whether a single IRT scale could encapsulate the important information about student proficiency in an area such as reading. In fact the BIB spiraling method was developed and applied to the 1983–1984 NAEP assessment precisely to address this question.

Rebecca Zwick (1987a, b) addressed this issue. Three methods were applied to the 1984 reading data: principal components analysis, full-information factor analysis (Bock et al. 1988), and a test of unidimensionality, conditional independence, and monotonicity based on contingency tables (Rosenbaum 1984). Results were consistent with the assumption of unidimensionality. A complicating factor in these analyses was the structure of the data that resulted from NAEP's BIB design. A simulation was conducted to investigate the impact of using the BIB-spiraled data in dimensionality analyses. Results from the simulated BIB data were similar to those from the complete data. The *Psychometrika* paper (Zwick 1987b), which describes some unique features of the correlation matrix of dichotomous Guttman items, was a spin-off of the NAEP research. Additional studies of dimensionality were performed by Carlson and Jirele (1992) and Carlson (1993).

Dimensionality has taken on increased importance as new uses are proposed for large-scale assessment data. Future survey design and analysis methods are evolving over time to address dimensionality as well as new factors that may affect the interpretation of assessment results. Some important factors are the need to ensure that the psychometric models incorporate developments in theories of how students learn, how changes in assessment frameworks affect performance, and how changes in the use of technology and integrated tasks affect results. Addressing these factors will require new psychometric models. These models will need to take into account specified relationships between tasks and underlying content domains, the cognitive processes required to solve these tasks, and the multilevel structure of the assessment sample. These models may also require development and evaluation of alternative estimation methods. Continuing efforts to further develop these methodologies include a recent methodological research project that is being conducted by ETS researchers Frank Rijmen and Matthias von Davier and is funded by the U.S. Department of Education's Institute of Education Sciences. This effort, through the application of a combination of general latent variable model frameworks (Rijmen et al. 2003; von Davier 2010) with new estimation methods based on stochastic (von Davier and Sinharay 2007, 2010) as well as a graphical model framework approach (Rijmen 2011), will offer a contribution to the research community that applies to NAEP as well as to other survey assessments.

The 1986 assessment produced unacceptable results, which have been referred to as the *reading anomaly*. The average score for 12th grade students fell by an estimated 2 years of growth, which could not have happened in the 2 years since the last assessment. The eighth grade students showed no decline, and the fourth graders showed a slight decline. This reading anomaly brought about a detailed exploration of possible explanations. Although a single factor was not isolated, it was concluded that many small changes produced the results. The results were published in a book by Beaton and Zwick (1990), who introduced the maxim "If you want to measure change, don't change the measure."

Further research was published by Zwick (1991). This paper summarized the key analyses described in the Beaton and Zwick reading anomaly report, focusing on the effects of changes in item position.

While confidence intervals for scaled scores are relatively straightforward, a substantial amount of research investigates confidence intervals for percentages (Brown et al. 2001; Oranje 2006a). NAEP utilizes an adjustment proposed by Satterthwaite (1941) to calculate effective degrees of freedom. However, Johnson and Rust (1993) detected through simulation that Satterthwaite's formula tends to underestimate effective degrees of freedom, which could cause the statistical tests to be too conservative. Qian (1998) conducted further simulation studies to support Johnson and Rust's conclusion. He also pointed out the instability associated with Satterthwaite's estimator.

8.3.9 NAEP's Technical Dissemination

An important contribution of ETS to large-scale group assessments is the way in which NAEP's substantive results and technology have been documented and distributed to the nation. This first part of this section will describe the many ways NAEP has been documented in publications. This will be followed by a discussion of the public-use data files and simple ways to perform secondary analyses using the NAEP data. The final section will present a description of some of the software available for advanced secondary analysts.

8.3.9.1 Documentation of NAEP Procedures and Results

ETS considered that communicating the details of the NAEP design and implementation was very important, and thus communication was promised in its winning proposal. This commitment led to a long series of publications, such as the following:

- *A New Design for a New Era* (Messick et al. 1983), which was a summary of the winning ETS NAEP proposal, including the many innovations that it planned to implement.
- The NAEP *Report Cards,* which give the results of NAEP assessments in different subject areas and different years. The first of these reports was *The Reading Report Card: Progress Toward Excellence in Our Schools, Trends in Reading over Four National Assessments, 1971–1984* (NAEP 1985).[14]
- NAEP Technical Reports,[15] which contain detailed information about sampling, assessment construction, administration, weighting, and psychometric methods. Beginning with the 2000 assessment, technical information has been published directly on the web.

[14] A full listing of such reports can be found at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031. These reports are complemented by press conferences.

[15] See http://nces.ed.gov/nationsreportcard/tdw/

- In 1992, two academic journal issues were dedicated to NAEP technology: *Journal of Educational Statistics*, Vol. 17, No. 2 (Summer, 1992) and *Journal of Educational Measurement*, Vol. 29, No. 2 (June, 1992).
- ETS has produced a series of reports to record technical contributions in NAEP. These scholarly works are included in the ETS Research publication series, peer reviewed by ETS staff and made available to the general public. A searchable database of such reports is available at http://search.ets.org/researcher/. Many of these reports are later published in professional journals.
- *The NAEP Primer*, written by Beaton and Gonzalez (1995) and updated extensively by Beaton et al. (2011).

8.3.9.2 NAEP's Secondary-Use Data and Web Tools

The NAEP staff has made extensive efforts to make its data available to secondary analysts. To encourage such uses, the NAEP design of 1983–1984 included public-use data files to make the data available. At that time, persons interested in secondary data analysis needed to receive a license from NCES before they were allowed to use the data files to investigate new educational policy issues. They could also check published statistics and explore alternative technologies. The public-use data files were designed to be used in commonly available statistical systems such as SPSS and SAS; in fact the choice of the plausible values technique was chosen in part over direct estimation methods to allow the data files tapes to use the rectangular format that was in general use at that time. Such files were produced for the 1984, 1986, and 1988 assessments.

The public-use data files did not bring about as much secondary analysis as hoped for. The complex technology introduced in NAEP, such as plausible values and replicate sampling weights, was intimidating. The data files contain very large numbers of students and school variables. To use the database properly required a considerable investment in comprehending the NAEP designs and analysis plans. The intellectual cost of using the public-use data files had discouraged many potential users.

In 1988, Congressional legislation authorized NAEP state assessments, beginning in 1990. Because of increased confidentiality concerns, the legislation precluded the issuing of public-use data files going forward. This action brought about a number of different approaches to data availability. The strict rules required by the Privacy Act (1974) made maintaining privacy more challenging. We will describe a few approaches to this problem in which ETS has played an important role.

Simple, Easily Available Products There are many potential users for the published NAEP graphs and tables and also for simple or complex variations on published outputs. Potential users include NAEP report writers and NAEP state coordinators, but also include educational policy makers, newspaper reporters, educational researchers, and interested members of the general public. To make the NAEP data available to such potential users, there was a need for computer programs

that were easy to use but employed the best available algorithms to help the users perform statistical analyses.

To respond to this need, ETS has developed and maintains web-based data tools for the purpose of analyzing large-scale assessment data. The foremost of these tools is the NAEP Data Explorer (NDE), whose principal developers at ETS were Alfred Rogers and Stephen Szyszkiewicz. NDE allows anyone with access to the Internet to navigate through the extensive, rich NAEP data archive and to produce results and reports that adhere to strict statistical, reporting, and technical standards. The user simply locates NDE on the web and, after electronically signing a user's agreement, is asked to select the data of interest: NAEP subject area; year(s) of assessment; states or other jurisdictions to be analyzed; and the correlates to be used in the analysis.[16]

NDE serves two sets of audiences: internal users (e.g., NAEP report writers and state coordinators) and the general public. NDE can be used by novice users and also contains many features appropriate for advanced users. Opening this data source to a much wider audience greatly increases the usefulness and transparency of NAEP. With a few clicks of a mouse, interested persons can effortlessly search a massive database, perform an analysis, and develop a report within a few minutes.

However, the NDE has its limitations. The NDE uses the full NAEP database and results from the NDE will be the same as those published by NAEP but, to ensure privacy, the NDE user is not allowed to view individual or school responses. The availability of statistical techniques is thus limited. NDE will refuse to compute statistics that might compromise individual responses, as might occur, for example, in a table in which the statistics in one or more cells are based on very small samples.

ETS has addressed making its data and techniques available through the *NAEP Primer* (Beaton et al. 2011). This publication for researchers provides much greater detail on how to access and analyze NAEP data, as well as an introduction to the available analysis tools and instruction on their use. A mini-sample of real data that have been approved for public use enables secondary analysts to familiarize themselves with the procedures before obtaining a license to a full data set. A NAEP-like data set is included for exploring the examples in the primer text.[17]

Full-Power, Licensed Products As mentioned above, using the NAEP database requires a substantial intellectual commitment. Keeping the NAEP subject areas, years, grades, and so forth straight is difficult and tedious. To assist users in the management of NAEP secondary-use data files, ETS developed the NAEP Data Toolkit. Alfred Rogers at ETS was the principal developer of the toolkit, which provides a data management application, NAEPEX, and procedures for performing two-way cross-tabulation and regression analysis. NAEPEX guides the user through the process of selecting samples and data variables of interest for analysis and

[16] The NDE is available free of charge at http://nces.ed.gov/nationsreportcard/naepdata/

[17] The primer is available at http://nces.ed.gov/nationsreportcard/researchcenter/datatools2.aspx

creates an extract data file or a set of SAS or SPSS control statements, which define the data of interest to the appropriate analysis system.[18]

Computational Analysis Tools Used for NAEP In addition to NAEPEX, ETS has developed a number of computer programs for more advanced users. These programs are intended to improve user access, operational ease, and computational efficiency in analyzing and reporting information drawn from the relatively large and complex large-scale assessment data sets. Continual development, enhancement, and documentation of applicable statistical methods and associated software tools are important and necessary. This is especially true given the ever increasing demand for—and scrutiny of—the surveys. Although initial large-scale assessment reports are rich and encyclopedic, there is great value in focused secondary analyses for interpretation, enhancing the value of the information, and formulation of policy. Diverse user audiences seeking to conduct additional analyses need to be confident in the methodologies, the computations, and in their ability to replicate, verify, and extend findings. The following presents a brief overview of several research-oriented computational analysis tools that have been developed and are available for both initial large-scale assessment operation and secondary research and analysis.

The methods and software required to perform direct estimation of group population parameters without introducing plausible values has developed substantially over the years. To analyze and report on the 1984 NAEP reading survey, ETS researchers and analysts developed the first operational version of the GROUP series of computer programs that estimate latent group effects. The GROUP series of programs is in continual development and advancement as evolving methods are incorporated. In addition to producing direct estimates of group differences, these programs may also produce plausible values based on Rubin's (1987) multiple imputations procedures for missing data. The output provides consistent estimates of population characteristics in filled-in data sets that enhance the ability to correctly perform secondary analyses with specialized software.

The separate programs in the GROUP series were later encapsulated into the DESI (Direct Estimation Software Interactive: ETS 2007; Gladkova et al. 2005) suite. DESI provides an intuitive graphical user interface (GUI) for ease of access and operation of the GROUP programs. The computational and statistical kernel of DESI can be applied to a broad range of problems, and the suite is now widely used in national and international large-scale assessments. WESVAR, developed at Westat, and the AM software program, developed at the American Institutes for Research (AIR) by Cohen (1998), also address direct estimation in general and are used primarily for analyzing data from complex samples, especially large-scale assessments such as NAEP. Descriptions and comparison of DESI and AM are found in papers by von Davier (2003) and Donoghue et al. (2006a). Sinharay and von Davier (2005) and von Davier and Sinharay (2007) discussed research around issues dealing with high performance statistical computing for large data sets found

[18] The NAEP Data Toolkit is available upon request from NAEP via http://nces.ed.gov/nationsreportcard/researchcenter/datatools2.aspx

in international assessments. Von Davier et al. (2006) presented an overview of large-scale assessment methodology and outlined steps for future extensions.

8.3.10 National Assessment Governing Board

The Elementary and Secondary Education act of 1988 authorized the national assessment governing board to set NAEP policies, schedules, and subject area assessment frameworks. This amendment made some important changes to the NAEP design. The main change was to allow assessment results to be reported by individual states so that the performance of students in various states could be compared. Such reporting was not permitted in previous assessments. This decision increased the usefulness and importance of NAEP. Reporting individual state results was introduced on a trial basis in 1990 and was approved as a permanent part of NAEP in 1996. Due to the success of individual state reporting, NAEP introduced separate reports for various urban school districts in 2002. These changes in NAEP reporting required vigilance to ensure that the new expanded assessments did not reduce the integrity of NAEP.

Several investigations were conducted to ensure the comparability and appropriateness of statistics over years and assessment type. Some of these are discussed in the sections below.

8.3.10.1 Comparability of State and National Estimate

At first, individual state reporting was done on a voluntary basis. The participating states needed large samples so that state subpopulations could be measured adequately. To maintain national population estimates, a sample of students from nonparticipating states was also collected. The participating and nonparticipating states' results were then merged with properly adjusted sampling weights. This separate sample for nonparticipating states became moot when all states participated as a result of the No Child Left Behind Act of 2002.

Two studies (Qian and Kaplan 2001; Qian et al. 2003) investigated the changes. The first described an analysis to ensure quality control of the combined national and state data. The second described the analyses directed at three main issues relevant to combining NAEP samples:

- Possible discrepancies in results between the combined sample and the current national sample.
- The effects of combined samples on the results of significance tests in comparisons, such as comparisons for reporting groups within the year and trend comparisons across years.
- The necessity of poststratification to adjust sample strata population estimates to the population values used in sample selection.

The findings of these studies showed that the combined samples will provide point estimates of population parameters similar to those from the national samples. Few substantial differences existed between combined and national estimates. In addition, the standard errors were smaller in the combined samples. With combined samples, there was a greater number of statistically significant differences in subpopulation comparisons within and across assessment years. The analysis also showed little difference between the results of nonpoststratified combined samples and those of poststratified combined samples.

8.3.10.2 Full Population Estimation

The publication of NAEP results for individual states allowed for comparisons of student performance. When more than one year was assessed in a subject area, estimation of trends in that area is possible. Trend comparisons are made difficult, since the published statistics are affected not only by the proficiency of students but also by the differences in the sizes of the subpopulations that are assessed. Early state trend results tended to show that states that excluded a larger percentage of students tended to have larger increases in reported average performance. This finding led to the search for full population estimates.

Although NAEP might like to estimate the proficiency of all students within an assessed grade, doing so is impractical. NAEP measurement tools cannot accurately measure the proficiency of some students with disabilities or students who are English language learners. While accommodations are made to include students with disabilities, such as allowing extra assessment time or use of braille booklets, some students are excluded. Despite strict rules for inclusion in NAEP, state regulations and practices vary somewhat and thus affect the comparability of state results.

To address this issue, Beaton (2000) suggested using a full population median, which Paul Holland renamed *bedian*. The bedian assumes only that the excluded students would do less well than the median of the full student population, and adjusts the included student median accordingly. McLaughlin (2000, 2005) proposed a regression approach by imputing excluded students' proficiencies from other available data. McLaughlin's work was further developed by Braun et al. (2008).

The National Institute of Statistical Sciences held a workshop on July 10–12, 2000, titled *NAEP Inclusion Strategies*. This workshop focused on comparing the full population statistics proposed by Beaton and McLaughlin. Included in its report is a detailed comparison by Holland (2000) titled "Notes on Beaton's and McLaughlin's Proposals."

8.3.11 Mapping State Standards Onto NAEP

The No Child Left Behind Act of 2002 required all states to set performance standards in reading and mathematics for Grades 3–8 and also for at least one grade in high school. The act, however, left to states the responsibility of determining the curriculum, selecting the assessments, and setting challenging academic standards. The result was that, in a particular grade, a standard such as *proficient* was reached by substantially different proportions of students in different states.

To understand the differences in state standards, ETS continued methodological development of an approach originally proposed by McLaughlin (1998) for making useful comparisons among state standards. It is assumed that the state assessments and NAEP assessment reflect similar content and have comparable structures, although they differ in test and item formats as well as standard-setting procedures. The Braun and Qian (2007) modifications involved (a) a shift from a school-based to a student-based strategy for estimating NAEP equivalent to a state standard, and (b) the derivation of a more refined estimate of the variance of NAEP parameter estimates by taking into account the NAEP design in the calculation of sampling error and by obtaining an estimate of the contribution of measurement error.

Braun and Qian applied the new methodology to four sets of data: (a) Year 2000 state mathematics tests and the NAEP 2000 mathematics assessments for Grades 4 and 8, and (b) Year 2002 state reading tests and the NAEP 2002 reading assessments for Grades 4 and 8. The study found that for both mathematics and reading, there is a strong negative linear relationship across states between the proportions meeting the standard and the apparent stringency of the standard as indicated by its NAEP equivalent. The study also found that the location of the NAEP score equivalent of a state's proficiency standard is not simply a function of the placement of the state's standard on the state's own test score scale. Rather, it also depends on the curriculum delivered to students across the state and the test's coverage of that curriculum with respect to both breadth and depth, as well as the relationship of both to the NAEP framework and the NAEP assessment administered to students. Thus, the variation among states' NAEP equivalent scores reflects the interaction of multiple factors, which can complicate interpretation of the results.

8.3.11.1 Testing Model Fit

IRT technology assumes that a student's response to an assessment item is dependent upon the students' ability, the item parameters of a known mathematical model, and an error term. The question arises as to how well the actual assessment data fit the assumed model. This question is particularly important in international assessments and also in any assessment where test items are translated into different languages. It is possible that the IRT model may fit well in one language but not well in another. For this reason, ETS applied an innovative model-fitting analysis for

comparing Puerto Rican students with mainland students. The Puerto Rican students responded to NAEP questions that were translated into Spanish.

The method for analyzing model fit was suggested by Albert Beaton (2003). The model was explored by Kelvin Gregory when he was at ETS. John Donoghue suggested using standardized errors in the comparison process. The method requires that the data set from an assessment has been analyzed using IRT and its results are available. Using the estimated student abilities and item parameters, a large number (e.g., 1000) of randomly equivalent data sets are created under the assumption of local independence. Statistics from the actual sample are then compared to the distribution of statistics from the randomly equivalent data sets. Large differences between the actual and randomly equivalent statistics indicate misfit. This approach indicates the existence of items or persons that do not respond as expected by the IRT model.

Additional research and procedures for assessing the fit of latent regression models was discussed by Sinharay et al. (2010). Using an operational NAEP data set, they suggested and applied a simulation-based model-fit procedure that investigated whether the latent regression model adequately predicted basic statistical summaries.

8.3.11.2 Aspirational Performance Standards

The National Assessment Governing Board decided to create achievement levels that were intended as goals for student performance. The levels were for *basic*, *proficient*, and *advanced*. Although ETS staff did not have a hand in implementing these levels, the standard-setting procedure of ETS researcher William Angoff (1971) was used in the early stages of the standard setting.

8.3.12 Other ETS Contributions

The ETS research staff continued to pursue technical improvements in NAEP under the auspices of the governing board, including those discussed in the following sections.

8.3.12.1 Rater Reliability in NAEP

Donoghue et al. (2006b) addressed important issues in rater reliability and the potential applicability of rater effects models for NAEP. In addition to a detailed literature review of statistics used to monitor and evaluate within- and across-year rater reliability, they proposed several alternative statistics. They also extensively discussed IRT-based rater-effect approaches to modeling rater leniency, and

provided several novel developments by applying signal detection theory in these models.

8.3.12.2 Computer-Based Assessment in NAEP

A key step towards computer-based testing in NAEP was a series of innovative studies in writing, mathematics, and critical reasoning in science and in technology-rich environments. The 2011 writing assessment was the first to be fully computer-based. Taking advantage of digital technologies enabled tasks to be delivered in audio and video multimedia formats. Development and administration of computer-delivered interactive computer tasks (ICTs) for the 2009 science assessment enabled measurement of science knowledge, processes, and skills that are not measurable in other modes. A mathematics online study in 2001 (Bennett et al. 2008) used both automated scoring and automatic item generation principles to assess mathematics for fourth and eighth graders on computers. This study also investigated the use of adaptive testing principles in the NAEP context. As of this writing, a technology and engineering literacy assessment is being piloted that assesses literacy as the capacity to use, understand, and evaluate technology, as well as to understand technological principles and strategies needed to develop solutions and achieve goals. The assessment is completely computer-based and engages students through the use of multimedia presentations and interactive simulations.

8.3.12.3 International Effects

The ETS methodology for group assessments has quickly spread around the world. At least seven major international studies have used or adapted the technology:

- School-based assessments
- The International Assessment of Educational Progress (IAEP)
- Trends in Mathematics and Science Study (TIMSS)
- Progress in International Reading Literacy Study (PIRLS)
- The Program for International Student Assessment (PISA 2015)
- Household-Based Adult Literacy Assessments
- The International Adult Literacy Study (IALS)
- The Adult Literacy and Life Skills Survey (ALL)
- The OECD Survey of Adult Skills. Also known as the Programme for the International Assessment of Adult Competencies (PIAAC)

In five of these studies (IAEP, PISA 2015, IALS, ALL, and PIAAC), ETS was directly involved in a leadership role and made significant methodological contributions. Two of the studies (TIMSS and PIRLS) have used ETS software directly under license with ETS and have received ETS scale validation services. These international assessments, including ETS's role and contributions, are described briefly below.

The existence of so many assessments brought about attempts to compare or link somewhat different tests. For example, comparing the IAEP test (Beaton and Gonzalez 1993) or linking the TIMSS test to NAEP tests might allow American students to be compared to students in foreign countries. ETS has carefully investigated the issues in linking and organized a special conference to address it. The conference produced a book outlining the problems and potential solutions (Dorans et al. 2007).

The IAEP assessments were conducted under the auspices of ETS and the UK's National Foundation for Educational Research, and funded by the National Science Foundation and NCES. In the middle of the 1980s there was concern about the start-up and reporting times of previously existing international assessments. In order to address these concerns, two assessments were conducted: IAEP1 in 1988 and IAEP2 in 1991. Archie Lapointe was the ETS director of these studies. Six countries were assessed in IAEP1. In IAEP2, students aged 9 and 13 from about 20 countries were tested in math, science, and geography. ETS applied the NAEP technology to these international assessments. These ventures showed that comprehensive assessments could be designed and completed quickly while maintaining rigorous standards. The results of the first IAEP are documented in a report titled *A World of Differences* (Lapointe et al. 1989). The IAEP methodologies are described in the *IAEP Technical Report* (1992).

The TIMSS assessments are conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). Conducted every 4 years since 1995, TIMSS assesses international trends in mathematics and science achievement at the fourth and eighth grades in more than 40 countries. For TIMSS, the ETS technology was adapted for the Rasch model by the Australian Council for Educational Research. The methodology used in these assessments was described in a TIMSS technical report (Martin and Kelly 1996).

The PIRLS assessments are also conducted under the auspices of the IEA. PIRLS is an assessment of reading comprehension that has been monitoring trends in student achievement at 5-year intervals in more than 50 countries around the world since 2001. PIRLS was described by Mullis et al. (2003).

The International Adult Literacy Survey (IALS), the world's first internationally comparative survey of adult skills, was administered in 22 countries in three waves of data collection between 1994 and 1998. The IALS study was developed by Statistics Canada and ETS in collaboration with participating national governments. The origins of the international adult literacy assessment program lie in the pioneering efforts employed in United States national studies that combined advances in large-scale assessment with household survey methodology. Among the national studies were the Young Adult Literacy Survey (Kirsch and Jungeblut 1986) undertaken by the NAEP program, and the National Adult Literacy Survey (described by Kirsch and ETS colleagues Norris**,** O'Reilly**,** Campbell**,** & Jenkins; Kirsch et al. 2000) conducted in 1992 by NCES.

ALL, designed and analyzed by ETS, continued to build on the foundation of IALS and earlier studies of adult literacy, and was conducted in 10 countries between 2003 and 2008 (Statistics Canada and OECD 2005).

The PIAAC study is an OECD Survey of Adult Skills conducted in 33 countries beginning in 2011. It measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper. The ETS Global Assessment Center, under the directorship of Irwin Kirsch, led the International Consortium and was responsible for the assessment's psychometric design, its analysis, and the development of cognitive assessment domains targeting skills in literacy, numeracy, and problem solving in technology-rich environments. ETS also coordinated development of the technology platform that brought the assessment to more than 160,000 adults, ages 16—65, in more than 30 language versions. The 2011 PIAAC survey broke new ground in international comparative assessment by being the first such instrument developed for computer-based delivery; the first to use multistage adaptive testing; the first to incorporate the use of computer-generated log file data in scoring and scaling; and the first to measure a set of reading components in more than 30 languages. The first PIAAC survey results were presented in an OECD publication (OECD 2013).

The PISA international study under the auspices of the OECD was launched in 1997. It aims to evaluate education systems worldwide every 3 years by assessing 15-year-olds' competencies in three key subjects: reading, mathematics, and science. To date, over 70 countries and economies have participated in PISA. For the sixth cycle of PISA in 2015, ETS is responsible for the design, delivery platform development, and analysis. To accomplish the new, complex assessment design, ETS Global continues to build on and expand the assessment methodologies it developed for PIAAC.

Kirsch et al. (Chap. 9, this volume) present a comprehensive history of Educational Testing Service's 25-year span of work in large-scale literacy assessments and resulting contributions to assessment methodology, innovative reporting, procedures, and policy information that "will lay the foundation for the new assessments yet to come."

In 2007, the Research and Development Division at ETS collaborated with the IEA Data Processing and Research Center to establish the IEA-ETS Research Institute (IERI). IERI publishes a SpringerOpen journal, *Large-Scale Assessments in Education*, which delivers state-of-the-art information on comparative international group score assessments. This IERI journal focuses on improving the science of large-scale assessments. A number of articles published in the IERI series present current research activities dealing with topics discussed in this paper, and also with issues surrounding the large-scale international assessments addressed here (TIMSS, PIRLS, PISA, IALS, ALL, and PIAAC).

In 2013, nine members of ETS's Research and Development division and two former ETSers contributed to a new handbook on international large-scale assessment (Rutkowski et al. 2014).

8.3.12.4 ETS Contributions to International Assessments

The ETS has also contributed to a number of international assessments in other ways, including the following:

- *GROUP Software*. GROUP software has been an important contribution of ETS to international assessments. This software gives many options for estimating the parameters of latent regression models, such as those used in national and international assessments. ETS offers licenses for the use of this software and consulting services as well. The software is described elsewhere in this paper and further described by Rogers et al. (2006).
- *International Data Explorer.* The NDE software has been adapted for international usage. The NDE allows a secondary researcher to create and manipulate tables from an assessment. ETS leveraged the NDE web-based technology infrastructure to produce the PIAAC Data Explorer (for international adult literacy surveys), as well as an International Data Explorer that reports on trends for PIRLS, TIMSS, and PISA data. The tools allow users to look up data according to survey, proficiency scale, country, and a variety of background variables, such as education level, demographics, language background, and labor force experiences. By selecting and organizing relevant information, stakeholders can use the large-scale data to answer questions of importance to them.
- *International linking.* Linking group assessments has taken on increased importance as new uses are proposed for large-scale assessment data. In addition to being linked to various state assessments, NAEP has been linked to TIMSS and PISA in order to estimate how well American students compare to students in other countries. In these cases, the tests being compared are designed to measure different—perhaps slightly different—student proficiencies. The question becomes whether or not the accuracy of a linking process is adequate for its proposed uses.

There is a wealth of literature on attempts at statistically linking national and international large-scale surveys to each other (Beaton and Gonzalez 1993; Johnson et al. 2003; Johnson and Siegendorf 1998; Pashley and Phillips 1993), as well as to state assessments (Braun and Qian 2007; McLaughlin 1998; Phillips 2007). Much of this work is based on concepts and methods of linking advocated by Mislevy (1992) and Linn (1993). In 2005, an ETS-sponsored conference focused on the general issue of score linking. The book that resulted from this conference (Dorans et al. 2007) examines the different types of linking both from theoretical and practical perspectives, and emphasizes the importance of both. It includes topics dealing with linking group assessments (such as NAEP and TIMSS). It also addresses mapping state or country standards to the NAEP scale.

There is an associated set of literature with arguments for and against the appropriateness of such mappings, and innovative attempts to circumvent some of the difficulties (Braun and Holland 1982; Linn and Kiplinger 1995; Thissen 2007; Wainer 1993). Past efforts to link large-scale assessments have met with varied levels of success. This called for continuing research to deal with problems such as

linking instability related to differences in test content, format, difficulty, measurement precision, administration conditions, and valid use. Current linking studies draw on this research and experience to ameliorate linking problems. For example, the current 2011 NAEP-TIMSS linking study is intended to improve on previous attempts to link these two assessments by administering NAEP and TIMSS booklets at the same time under the same testing conditions, and using actual state TIMSS results in eight states to validate the predicted TIMSS average scores.

8.3.13 NAEP ETS Contributions

Large-scale group assessments lean heavily on the technology of other areas such as statistics, psychometrics, and computer science. ETS researchers have also contributed to the technology of these areas. This section describes a few innovations that are related to other areas as well as large-scale group assessments.

8.3.13.1 The FORTRAN IV Statistical System (F4STAT)

Although the development of F4STAT began in 1964, before ETS was involved in large-scale group assessments,[19] it quickly became the computation engine that made flexible, efficient data analysis possible. Statistical systems of the early 60s were quite limited and not generally available. Typically, they copied punch card systems that were used on earlier computers. Modern systems such as SAS, SPSS, and Stata were a long way off.

ETS had ordered an IBM 7040 computer for delivery in 1965, and it needed a new system that would handle the diverse needs of its research staff. For this reason, the organization decided to build its own statistical system, F4STAT (Beaton 1973b). Realizing that parameter-driven programs could not match the flexibility of available compilers, the decision was made to use the Fortran IV compiler as the driving force and then develop statistical modules as subroutines. Based on the statistical calculus operators defined by Beaton (1964), the F4STAT system was designed to be modular, general, and easily expandable as new analytic methods were conceived. Of note is that the Beaton operators are extensively cited and referenced throughout statistical computation literature (Dempster 1969; Milton and Nelder 1969), and that these operators or their variants are used in commercial statistical systems, such as SAS and SPSS (Goodnight 1979). Through incorporation of a modern integrated development environment (IDE), F4STAT continues to provide the computational foundation for ETS's large-scale assessment data analysis systems. This continual, technology-driven evolution is important for ETS researchers

[19] Albert Beaton, William Van Hassel, and John Barone implemented the early ETS F4STAT system. Ongoing development continued under Barone. Alfred Rogers is the current technical leader.

to respond to the ever increasing scope and complexity of large-scale and longitudinal surveys and assessments.

8.3.13.2 Fitting Robust Regressions Using Power Series

Many data analyses and, in particular large-scale group assessments, rely heavily on minimizing squared residuals, which overemphasizes the larger residuals. Extreme outliers may completely dominate an analysis. Robust regression methods have been developed to provide an alternative to least squares regression by detecting and minimizing the effect of deviant observations. The primary purpose of robust regression analysis is to fit a model that represents the information in the majority of the data. Outliers are identified and may be investigated separately.

As a result, the issue of fitting power series became an important issue at this time. Beaton and Tukey (1974) wrote a paper on this subject, which was awarded the Wilcoxon Award for the best paper in *Technometrics* in that year. The paper led to a method of computing regression analyses using least absolute value or minimax criteria instead of least squares. For more on this subject, see Holland and Welsch (1977), who reviewed a number of different computational approaches for robust linear regression and focused on iteratively reweighted least-squares (IRLS). Huber (1981, 1996) presented a well-organized overview of robust statistical methods.

8.3.13.3 Computational Error in Regression Analysis

An article by Longley (1967) brought about concern about the accuracy of regression programs. He found large discrepancies among the results of various regression programs. Although ETS software was not examined, the large differences were problematic for any data analyst. If regression programs were inconsistent, large-scale group studies would be suspect.

To investigate this problem, Beaton et al. (1976) looked carefully at the Longley data. The data were taken from economic reports and rounded to thousands, millions, or whatever depending on the variable. The various variables were highly collinear. To estimate the effect of rounding, they added a random uniform number to each datum in the Longley analysis. These random numbers had a mean of zero and a range of -.5 to +.5 after the last published digit. One thousand such data sets were produced, and each set would round to the published data.

The result was surprising. The effect of these random digits substantially affected the regression results more than the differences among various programs. In fact, the "highly accurate" results—computed by Longley to hundreds of places—were not even at the center of the distribution of the 1,000 regression results. The result was clear: increasing the precision of calculations with near-collinear data is not worth the effort, the "true" values are not calculable from the given data.

This finding points out that a greater source of inaccuracy may be the data themselves. Cases such as this, where slight variations in the original data cause large

variations in the results, suggest further investigation is warranted before accepting the results. The cited ETS paper also suggests a ridge regression statistic to estimate the seriousness of collinearity problems.

8.3.13.4 Interpreting Least Squares

Regression analysis is an important tool for data analysis in most large- and small-scale studies. Generalizations from an analysis are based on assumptions about the population from which the data are sampled. In many cases, the assumptions are not met. For example, EOS had a complex sample and a 65% participation rate and therefore did not meet the assumptions for regression analysis. Small studies, such as those that take the data from an almanac, seldom meet the required assumptions. The purpose of this paper is to examine what can be stated without making any sampling assumptions.

Let us first describe what a typical regression analysis involves. Linear regression assumes a model such as $y = X\beta+\varepsilon$, where y is the phenomenon being studied, X represents explanatory variables, β is the set of parameters to be estimated, and ε is the residual. In practice, where N is the number of observations ($i = 1,2,\ldots,N$) and M ($j = 0,1,\ldots,M$) is the number of explanatory variables, y is an Nth order vector, X is an N x M matrix, β is an Mth order vector, and ε is an Nth order vector. The values $x_{i0} = 1$ and β_0 = the intercept. The values in y and X are assumed to be known. The values in ε are assumed to be independently distributed from a normal distribution with mean of 0 and variance of σ^2. Regression programs compute b, the least squares estimate of β, s^2 the estimate of σ^2, and e, the estimate of ε. Under the assumptions, regression creates a t-test for each regression coefficient in b, testing the hypotheses that $\beta_j = 0$. A two-tailed probability statistic p_j is computed to indicate the probability of obtaining a b_j if the true value is zero. A regression analysis often includes an F test that tests the hypothesis that all regression coefficients (excluding the intercept) are equal to zero.

The question addressed here is what we can say about the regression results if we do not assume that the error terms are randomly distributed. Here, we look at the regression analysis as a way of summarizing the relationship between the y and X variables. The regression coefficients are the summary. We expect a good summary to allow us to approximate the values of y using the X variables and their regression coefficients. The question then becomes: How well does the model fit?

Obviously, a good fit implies that the errors are small, near zero. Small errors should not have a substantial effect on the data summary, that is, the regression coefficients. The effect of the error can be evaluated by permuting the errors and then computing the regression coefficients using the permuted data. There are $N!$ ways to permute the errors. Paul Holland suggested flipping the signs of the errors. There are 2^N possible ways to flip the error signs. Altogether, there are $N!2^N$ possible signed permutations, which is a very large number. For example, 10 observations generate 3,628,800 × 1,024 = 3,715,891,200 possible signed permutations. We will

denote each signed permutations as e_k (k = 1,2,…, $2^N N!$,), $y_k = X\beta + e_{k,}$ and the corresponding regression coefficient as b_k with elements b_{jk}.

Fortunately, we do not need to compute these signed permutations to describe the model fit. Beaton (1981) has shown that the distribution of sign permuted regression coefficients rapidly approaches a normal distribution as the number of observations increases. The mean of the distribution is the original regression coefficient, and the standard deviation is approximately the same as the standard error in regression programs.

The model fit can be assessed from the p values computed in a regular regression analysis:

- The probability statistic p_j for an individual regression coefficient can be interpreted as the proportion of signed and permuted regression coefficients b_{jk} that are further away from b_j than the point where the b_{jk} have different signs.
- Since the distribution is symmetric, $.5p_j$ can be interpreted as the percentage of the b_{jk} that have different signs from b_j.
- The overall P statistic can be interpreted as the percentage of b_k that is as far from b as the point where all b_k have a different sign.
- Other fit criteria are possible, such as computing the number of b_{jk} that differ in the first decimal place.

In summary, the model fit is measured by comparing the sizes of the errors to their effect on the regression coefficients. The errors are not assumed to come from any outside randomization process. This interpretation is appropriate for any conforming data set. The ability to extrapolate to other similar data sets is lost by the failure to assume a randomization.

8.3.14 Impact on Policy—Publications Based on Large-Scale Assessment Findings

Messick (1986) described analytic techniques that provide the mechanisms for inspecting, transforming, and modeling large-scale assessment data with the goals of providing useful information, suggesting conclusions, and supporting decision making and policy research. In this publication, Messick eloquently espoused the enormous potential of large-scale educational assessment as effective policy research and examined critical features associated with transforming large-scale educational assessment into effective policy research. He stated that

> In policy research it is not sufficient simply to document the direction of change, which often may only signal the presence of a problem while offering little guidance for problem solution. One must also conceptualize and empirically evaluate the nature of the change and its contributing factors as a guide for rational decision making.

Among the critical features that he deemed necessary are the capacity to provide measures that are commensurable across time periods and demographic groups,

correlational evidence to support construct interpretations, and multiple measures of diverse background and program factors to illuminate context effects and treatment or process differences. Combining these features with analytical methods and interpretative strategies that make provision for exploration of multiple perspectives can yield relevant, actionable policy alternatives. Messick noted that settling for less than full examination of plausible alternatives due to pressures of timeliness and limited funding can be, ironically, at the cost of timeliness.

With the above in mind, we refer the reader to the NCES and ETS websites to access the links to a considerable collection of large-scale assessment publications and data resources. Also, Coley, Goertz, and Wilder (Chap. 12, this volume) provide additional policy research insight.

Appendix: NAEP Estimation Procedures

The NAEP estimation procedures start with the assumption that the proficiency of a student in an assessment area can be estimated from a student's responses to the assessment items that the student received. The psychometric model is a latent regression consisting of four types of variables:

- Student proficiency
- Student item responses
- Conditioning variables
- Error variables

The true proficiency of a student is unobservable and thus unknown. The student item responses are known, since they are collected in an assessment. Also known are the conditioning variables that are collected for reporting (e.g., demographics) or may be otherwise considered related to student proficiency. The error variable is the difference between the actual student proficiency and its estimate from the psychometric model and is thus unknown.

The purpose of this appendix is to present the many ways in which ETS researchers have addressed the estimation problem and continue to look for more precise and efficient ways of using the model. Estimating the parameters of the model requires three steps:

1. Scaling
2. Conditioning
3. Variance estimation

Scaling processes the item-response statistics to develop estimates of student proficiency. Conditioning adjusts the proficiency estimates in order to improve their accuracy and reduce possible biases. Conditioning is an iterative process using the estimation–maximization (EM) algorithm (Dempster et al. 1977) that leads to maximum likelihood estimates. Variance estimation is the process by which the error in

the parameter estimates is itself estimated. Both sampling and measurement error are examined.

The next section presents some background on the original application of this model. This is followed by separate sections on advances in scaling, conditioning, and variance estimation. Finally, a number of alternate models proposed by others are evaluated and discussed.

The presentation here is not intended to be highly technical. A thorough discussion of these topics is available in a section of the *Handbook of Statistics* titled "Marginal Estimation of Population Characteristics: Recent Developments and Future Directions" (von Davier et al. 2006).

The Early NAEP Estimation Process

NAEP procedures proposed by ETS were conceptually straightforward: the item responses are used to estimate student proficiency, and then the student estimates are summarized by gender, racial/ethnic groupings, and other factors of educational importance. The accuracy of the group statistics would be estimated using sampling weights and the jackknife method which would take into account the complex NAEP sample. The 3PL IRT model was to be used as described in Lord and Novick (1968).

This approach was first used in the 1983–1984 NAEP assessment of reading and writing proficiency. The proposed IRT methodology of that time was quite limited: it handled only multiple-choice items that could be scored either right or wrong. It also could not make any finite estimates for students who answered all items correctly or scored below the chance level. Since the writing assessment had graded-response questions, the standard IRT programs did not work, so the ARM was developed by Beaton and Johnson (1990). The ARM was later replaced by the PARSCALE program (Muraki and Bock 1997).

However, the straightforward approach to reading quickly ran into difficulties. The decision had been made to BIB spiral the reading and writing items, with the result that many students were assigned too few items to produce an acceptable estimate of their reading proficiency. Moreover, different racial/ethnic groupings had substantially different patterns of inestimable proficiencies, which would bias any results. Standard statistical methods did not offer any solution.

Fortunately, Mislevy had the insight that NAEP did not need individual student proficiency estimates; it needed only estimates of select populations and subpopulations. This led to the use of marginal maximum likelihood methods through the BILOG program (Mislevy and Bock 1982). The BILOG program could estimate group performance directly, but an alternative approach was taken in order to make the NAEP database useful to secondary researchers. BILOG did not develop acceptable individual proficiency estimates but did produce a posterior distribution for each student that indicated the likelihood of possible estimates. From these distributions, five plausible values were randomly selected. Using these plausible values

made data analysis more cumbersome but produced a data set that could be used in most available statistical systems.

The adaptation and application of this latent regression model was used to produce the NAEP 1983–1984 Reading Report Card, which has served as a model for many subsequent reports. More details on the first application of the NAEP estimation procedures were described by Beaton (1987) and Mislevy et al. (1992).

Scaling

IRT is the basic component of NAEP scaling. As mentioned above, the IRT programs of the day were limited and needed to be generalized to address NAEP's future needs. There were a number of new applications, even in the early NAEP analyses:

- Vertical scales that linked students aged 9, 13, and 17.
- Across-year scaling to link the NAEP reading scales to the comparable assessments in the past.
- In 1986, subscales were introduced for the different subject areas. NAEP produced five subscales in mathematics. Overall mathematics proficiency was estimated using a composite of the subscales.
- In 1992, the generalized partial credit model was introduced to account for graded responses (polytomous items) such as those in the writing assessments (Muraki 1992; Muraki and Bock 1997).

Yamamoto and Mazzeo (1992) presented an overview of establishing the IRT-based common scale metric and illustrated the procedures used to perform these analyses for the 1990 NAEP mathematics assessment. Muraki et al. (2000) provided an overview of linking methods used in performance assessments, and discussed major issues and developments in linking performance assessments.

Conditioning

As mentioned, the NAEP reporting is focused on group scores. NAEP collected a large amount of demographic data, including student background information and school and teacher questionnaire data, which can be used to supplement the nonresponse due to BIB design and to improve the accuracy of group scores.

Mislevy (1984, 1985) has shown that maximum likelihood estimates of the parameters in the model can be obtained when the actual proficiencies are unknown using an EM algorithm.

The NAEP conditioning model employs both cognitive data and demographic data to construct a latent regression model. The implementation of the EM algorithm that is used in the estimation of the conditioning model leaves room for

possible improvements in accuracy and efficiency. In particular, there is a complex multidimensional integral that must be calculated, and there are many ways in which this can be done, each method embodied by a computer program which has been carefully investigated for advantages and disadvantages. These programs have been generically labeled as GROUP programs. The programs that have been used or are currently in use are as follows:

- BGROUP (Sinharay and von Davier 2005). This program is a modification of BILOG (Mislevy and Bock 1982) and uses numerical quadrature and direct integration. This is typically used when there are one or two scales being analyzed
- MGROUP (Mislevy and Sheehan 1987) uses a Monte Carlo method to draw random normal estimates from posterior distributions as input to each estimation step.
- NGROUP (Allen et al. 1996; Mislevy 1985) uses Bayesian normal theory. The requirement of the assumption of a normal distribution results in little use of this method.
- CGROUP (Thomas 1993) uses a Laplace approximation for the posterior means and variance. This method is used when more than two scales are analyzed.
- DGROUP (Rogers et al. 2006) is the current operational program that brings together the BGROUP and CGROUP methods on a single platform. This platform is designed to allow inclusion of other methods as they are developed and tested.

To make these programs available in a single package, ETS researchers Ted Blew, Andreas Oranje, Matthias von Davier, and Alfred Rogers developed a single program called DESI that allows a user to try the different latent regression programs.

The end result of these programs is a set of plausible values for each student. These are random draws from each student's posterior distribution, which gives the likelihood of a student having a particular proficiency score. The plausible value methodology was developed by Mislevy (1991) based on the ideas of Little and Rubin (1987, 2002) on multiple imputation. These plausible values are not appropriate for individual proficiency scores or decision making. In their 2009 paper, "What Are Plausible Values and Why Are They Useful*?*," von Davier et al. described how plausible values are applied to ensure that the uncertainty associated with measures of skills in large scale surveys is properly taken into account. In 1988, NCME gave its Award for Technical Contribution to Educational Measurement to ETS researchers Robert Mislevy, Albert Beaton, Eugene Johnson, and Kathleen Sheehan for the development of plausible values methodology in the NAEP.

The student plausible values are merged with their sampling weights to compute population and subpopulation statistical estimates, such as the average student proficiency of a subpopulation.

It should be noted that the AM method (Cohen 1998) estimates population parameters directly and is a viable alternative to the plausible-value method that ETS has chosen. The AM approach has been studied in depth by Donoghue et al. (2006a).

These methods were subsequently evaluated for application in future large-scale assessments (Li and Oranje 2006; Sinharay et al. 2010; Sinharay and von Davier 2005; von Davier and Sinharay 2007, 2010). Their analysis of a real NAEP data set provided some evidence of a misfit of the NAEP model. However, the magnitude of the misfit was small, which means that the misfit probably had no practical significance. Research into alternative approaches and emerging methods is continuing.

Variance Estimation

Error variance has two components: sampling error and measurement error. These components are considered to be independent and are summed to estimate total error variance.

Sampling Error

The NAEP samples are obtained through a multistage probability sampling design. Because of the similarity of students within schools and of the effects of nonresponse, observations made of different students cannot be assumed to be independent of each other. To account for the unequal probabilities of selection and to allow for adjustments for nonresponse, each student is assigned separate sampling weights. If these weights are not applied in the computation of the statistics of interest, the resulting estimates can be biased. Because of the effects of a complex sample design, the true sampling variability is usually larger than a simple random sampling. More detailed information is available in reports by Johnson and Rust (1992, 1993), Johnson and King (1987), and Hsieh et al. (2009).

The sampling error is estimated by the jackknife method (Quenouille 1956; Tukey 1958). The basic idea is to divide a national or state population, such as in-school eighth graders, into primary sampling units (PSUs) that are reasonably similar in composition. Two schools are selected at random from each PSU. The sampling error is estimated by computing as many error estimates as there are PSUs. Each of these replicates consists of all PSU data except for one, in which one school is randomly removed from the estimate and the other is weighted doubly. The methodology for NAEP was described, for example, by E. G. Johnson and Rust (1992), and von Davier et al. (2006), and a possible extension was discussed by Hsieh et al. (2009).

The sampling design has evolved as NAEP's needs have increased. Certain ethnic groups are oversampled to ensure that reasonably accurate estimations and sampling weights are developed to ensure appropriately estimated national and state samples.

Also, a number of studies have been conducted about the estimation of standard errors for NAEP statistics. Particularly, an application of the Binder methodology (see also Cohen and Jiang 2001) was evaluated (Li and Oranje 2007) and a

comparison with other methods was conducted (Oranje et al. 2009) showing that the Binder method under various conditions underperformed compared to sampling-based methods.

Finally, smaller studies were conducted on (a) the use of the coefficient of variation in NAEP (Oranje 2006b), which was discontinued as a result; (b) confidence intervals for NAEP (Oranje 2006a), which are now available in the NDE as a result; and (c) disclosure risk prevention (Oranje et al. 2007), which is currently a standard practice for NAEP.

Measurement Error

Measurement error is the difference between the estimated results and the "true" results that are not usually available. The plausible values represent the posterior distribution and can be used for estimating the amount of measurement error in statistical estimates such as a population mean or percentile. Five plausible values are computed for each student, and each is an estimate of the student's proficiency. If the five plausible values are close together, then the student is well measured; if the values differ substantially, the student is poorly measured. The variance of the plausible values over an entire population and subpopulation can be used to estimate the error variance. The general methodology was described by von Davier et al. (2009).

Researchers continue to explore alternative approaches to variance estimation for NAEP data. For example, Hsieh et al. (2009) explored a resampling-based approach to variance estimation that makes ability inferences based on replicate samples of the jackknife without using plausible values.

Alternative Psychometric Approaches

A number of modifications of the current NAEP methodology have been suggested in the literature. These evolved out of criticisms of (a) the complex nature of the NAEP model and (b) the approximations made at different stages of the NAEP estimation process. Several such suggestions are listed below:

- *Apply a group-specific variance term*. Thomas (2000) developed a version of the CGROUP program that allowed for a group-specific residual variance term instead of assuming a uniform term across all groups.
- *Apply seemingly unrelated regressions* (SUR; Greene 2002; Zellner 1962). Researchers von Davier and Yu (2003) explored this suggestion using a program called YGROUP and found that it generated slightly different results from CGROUP. Since YGROUP is faster, it may be used to produce better starting values for the CGROUP program.

- *Apply a stochastic EM method.* Researchers von Davier and Sinharay (2007) approximated the posterior expectation and variance of the examinees' proficiencies using importance sampling (e.g., Gelman et al. 2004). Their conclusion was that this method is a viable alternative to the MGROUP system but does not present any compelling reason for change.
- *Apply stochastic approximation.* A promising approach for estimation in the presence of high dimensional latent variables is stochastic approximation. Researchers von Davier and Sinharay (2010) applied this approach to the estimation of conditioning models and showed that the procedure can improve estimation in some cases.
- *Apply multilevel IRT using Markov chain Monte Carlo methods (MCMC).* M. S. Johnson and Jenkins (2004) suggested an MCMC estimation method (e.g., Gelman et al. 2004; Gilks et al. 1996) that can be adapted to combine the three steps (scaling, conditioning, and variance estimation) of the MGROUP program. This idea is similar to that proposed by Raudenbush and Bryk (2002). A maximum likelihood application of this model was implemented by Li et al. (2009) and extended to dealing with testlets by Wang et al. (2002).
- *Estimation using generalized least squares (GLS).* Researchers von Davier and Yon (2004) applied GLS methods to the conditioning model used in NAEP's MGROUP, employing an individual variance term derived from the IRT measurement model. This method eliminates some basic limitations of classical approaches to regression model estimation.
- *Other modifications.* Other important works on modification of the current NAEP methodology include those by Bock (2002) and Thomas (2002).

Possible Future Innovations

Random Effects Model

ETS developed and evaluated a random effects model for population characteristics estimation. This approach explicitly models between-school variability as a random effect to determine whether it is better aligned with the observed structure of NAEP data. It was determined that relatively small gains in estimation using this approach in NAEP were not sufficient to override the increase in computational complexity. However, this approach does appear to have potential for use in international assessments such as PISA and PIRLS.

Adaptive Numerical Quadrature

Use of adaptive numerical quadrature can improve estimation accuracy over using approximation methods in high-dimensional proficiency estimation. ETS researchers performed analytic studies (Antal and Oranje 2007; Haberman 2006) using

adaptive quadrature to study the benefit of increased precision through numerical integration for multiple dimensions. Algorithmic development and resulting evaluation of gains in precision are ongoing, as are feasibility studies for possible operational deployment in large-scale assessment estimation processes.

Antal and Oranje (2007) posited that the Gauss-Hermite rule enhanced with Cholesky decomposition and normal approximation of the response likelihood is a fast, precise, and reliable alternative for the numerical integration in NAEP and in IRT in general.

Using Hierarchical Models

In addition, several studies have been conducted about the use of hierarchical models to estimate latent regression effects that ultimately lead to proficiency estimates for many student groups of interest. Early work based on MCMC (Johnson and Jenkins 2004) was extended into an MLE environment, and various studies were conducted to evaluate applications of this model to NAEP (Li et al. 2009).

The NAEP latent regression model has been studied to understand better some boundary conditions under which the model performs well or not so well (Moran and Dresher 2007). Research into different approaches to model selection has been initiated (e.g., Gladkova and Oranje 2007). This is an ongoing project.

References

Allen, N. L., Johnson, E. J., Mislevy, R. J., & Thomas, N. (1996). Scaling procedures. In N. L. Allen, D. L. Kline, & C. A. Zelenak (Eds.), *The NAEP 1994 technical report* (pp. 247–266). Washington, DC: National Center for Education Statistics.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Antal, T., & Oranje, A. (2007). *Adaptive numerical integration for item response theory* (Research Report No. RR-07-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02048.x

Beall, G., & Ferris, J. (1971). *On discovering Youden rectangles with columns of treatments in cyclic order* (Research Bulletin No. RB-71-37). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1971.tb00611.x

Beaton, A. E. (1964). *The use of special matrix operators in statistical calculus* (Research Bulletin No. RB-64-51). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1964.tb00689.x

Beaton, Albert E. (1968). *Some considerations of technical problems in the Educational Opportunity Survey* (Research Memorandum No. RM-68-17). Princeton: Educational Testing Service.

Beaton, A. E. (1969). Scaling criterion of questionnaire items. *Socio–Economic Planning Sciences, 2*, 355–362. https://doi.org/10.1016/0038-0121(69)90030-5

Beaton, A. E. (1973a). *Commonality*. Retrieved from ERIC Database. (ED111829)

Beaton, A. E. (1973b). F4STAT statistical system. In W. J. Kennedy (Ed.), *Proceedings of the computer science and statistics: Seventh annual symposium of the interface* (pp. 279–282). Ames: Iowa State University Press.

Beaton, A. E. (1975). Ability scores. In F. T. Juster (Ed.), *Education, income, and human behavior* (pp. 427–430). New York: McGraw-Hill.

Beaton, A. E. (1981). *Interpreting least squares without sampling assumptions* (Research Report No. RR-81-38). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1981.tb01265.x

Beaton, A. E. (1987). *The NAEP 1983−84 technical report*. Washington, DC: National Center for Education Statistics.

Beaton, A. E. (2000). *Estimating the total population median.* Paper presented at the National Institute of Statistical Sciences workshop on NAEP inclusion strategies. Research Triangle Park: National Institute of Statistical Sciences.

Beaton, A. (2003). *A procedure for testing the fit of IRT models for special populations*. Unpublished manuscript.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191–204. https://doi.org/10.2307/1165169

Beaton, A. E., & Chromy, J. R. (2007). *Partitioning NAEP trend data*. Palo Alto: American Institutes for Research.

Beaton, A. E., & Gonzalez, E. J. (1993). *Comparing the NAEP trial state assessment results with the IAEP international results*. *Report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment*. Stanford: National Academy of Education.

Beaton, A. E., & Gonzalez, E. (1995). *NAEP primer*. Chestnut Hill: Boston College.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9–38. https://doi.org/10.2307/1164819

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band–spectroscopic data. *Technometrics, 16*, 147–185. https://doi.org/10.1080/00401706.1974.10489171

Beaton, A.E., & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985–86 reading anomaly* (NAEP Report No. 17–TR–21). Princeton: Educational Testing Service.

Beaton, A. E., Rubin, D. B., & Barone, J. L. (1976). The acceptability of regression solutions: Another look at computational accuracy. *Journal of the American Statistical Association, 71*, 158–168. https://doi.org/10.1080/01621459.1976.10481507

Beaton, A. E., Hilton, T. L., & Schrader, W. B. (1977). *Changes in the verbal abilities of high school seniors, college entrants, and SAT candidates between 1960 and 1972* (Research Bulletin No. RB-77-22). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1977.tb01147.x

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., … Jia, Y. (2011). *The NAEP primer* (NCES Report No. 2011–463). Washington, DC: National Center for Education Statistics.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*, 1–39.

Bock, R.D. (2002). *Issues and recommendations on NAEP data analysis*. Palo Alto: American Institutes for Research.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full–information item factor analysis. *Applied Psychological Measurement, 12,* 261–280. https://doi.org/10.1177/014662168801200305

Bowles, S., & Levin, H. M. (1968). The determinants of scholastic achievement: An appraisal of some recent evidence. *Journal of Human Resources, 3*, 3–24.

Braun, H. I., & Holland, P. W. (1982). Observed–score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_17

Braun, H., Zhang, J., & Vezzu, S. (2008). *Evaluating the effectiveness of a full-population estimation method* (Research Report No. RR-08-18). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02104.x

Bridgeman, B., Blumenthal, J. B., & Andrews, S. R. (1981). *Parent child development center: Final evaluation report*. Unpublished manuscript.

Brown, L. D., Cai, T., & DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science, 16*, 101–133. https://doi.org/10.1214/ss/1009213286

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods*. Newbury Park: Sage.

Cain, G., & Watts, H. W. (1968). The controversy about the Coleman report: Comment. *The Journal of Human Resources, 3*, 389–392. https://doi.org/10.2307/145110

Carlson, J. E. (1993, April). *Dimensionality of NAEP instruments that incorporate polytomously-scored items*. Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

Carlson, J. E., & Jirele, T. (1992, April). *Dimensionality of 1990 NAEP mathematics data*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Civil Rights Act, P.L. No. 88-352, 78 Stat. 241 (1964).

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28*, 345–360. https://doi.org/10.1177/001316446802800212

Cohen, J. D. (1998). *AM online help content—Preview*. Washington, DC: American Institutes for Research.

Cohen, J., & Jiang, T. (2001). *Direct estimation of latent distributions for large-scale assessments with application to the National Assessment of Educational Progress (NAEP)*. Washington, DC: American Institutes for Research.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U. S. Government Printing Office.

Dempster, A. P. (1969). *Elements of continuous multivariate analysis*. Reading: Addison–Wesley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Donoghue, J., Mazzeo, J., Li, D., & Johnson, M. (2006a). *Marginal estimation in NAEP: Current operational procedures and AM*. Unpublished manuscript.

Donoghue, J., McClellan, C. A., & Gladkova, L. (2006b). *Using rater effects models in NAEP*. Unpublished manuscript.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Gamoran, A., & Long, D. A. (2006). *Equality of educational opportunity: A 40-year retrospective*. (WCER Working Paper No. 2006-9). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton: Chapman and Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Gladkova, L., & Oranje, A. (2007, April). *Model selection for large scale assessments.* Paper presented at the meeting of the National Council of Measurement in Education, Chicago, IL.

Gladkova, L., Moran, R., Rogers, A., & Blew, T. (2005). Direct estimation software interactive (DESI) manual [Computer software manual]. Princeton: Educational Testing Service.

Goodnight, J. H. (1979). A tutorial on the SWEEP operator. *American Statistician, 33*, 149–158. https://doi.org/10.1080/00031305.1979.10482685

Greene, W. H. (2002). *Econometric analysis* (5th ed.). Upper Saddle River: Prentice Hall.

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (Research Report No. RR-06-29). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02035.x

Hilton, T. L. (1992). *Using national data bases in educational research*. Hillsdale: Erlbaum.

Holland, P. W. (2000). Notes on Beaton's and McLaughlin's proposals. In L. V. Jones & I. Olkin, *NAEP inclusion strategies: The report of a workshop at the National Institute of Statistical Sciences.* Unpublished manuscript.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics – Theory and Methods, A6*, 813–827. https://doi.org/10.1080/03610927708827533

Hsieh, C., Xu, X., & von Davier, M. (2009). Variance estimation for NAEP data using a resampling–based approach: An application of cognitive diagnostic models. *IERI Monograph Series: Issues and methodologies in large scale assessments, 2*, 161–173.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley. https://doi.org/10.1002/0471725250

Huber, P. J. (1996). *Robust statistical procedures* (2nd ed.). Philadelphia: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970036

International Assessment of Educational Progress. (1992). *IAEP technical report*. Princeton: Educational Testing Service.

Johnson, M. S., & Jenkins, F. (2004). *A Bayesian hierarchical model for large–scale educational surveys: An application to the National Assessment of Educational Progress* (Research Report No. RR-04-38). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01965.x

Johnson, E. G., & King, B. F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics, 3*, 235–250. https://doi.org/10.1002/j.2330-8516.1987.tb00210.x

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190. https://doi.org/10.2307/1165168

Johnson, E. G., & Rust, K. F. (1993). Effective degrees of freedom for variance estimates from a complex sample survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 863–866). Alexandria, VA: American Statistical Association.

Johnson, E. G., & Siegendorf, A. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): Eighth–grade results* (NCES Report No. 98–500). Washington, DC: National Center for Education Statistics.

Johnson, E., Cohen, J., Chen, W. H., Jiang, T., & Zhang, Y. (2003). *2000 NAEP-1999 TIMSS linking report* (NCES Publication No. 2005–01). Washington, DC: National Center for Education Statistics.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The Nation's Report Card: Evolutions and perspectives*. Bloomington: Phi Delta Kappa Educational Foundation.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults* (NAEP Report No. 16-PL-01). Princeton: National Assessment of Educational Progress.

Kirsch, I., Yamamoto, K., Norris, N., Rock, D., Jungeblut, A., O'Reilly, P., … Baldi, S. (2000). *Technical report and data files user's manual For the 1992 National Adult Literacy Survey*. (NCES Report No. 2001457). U.S. Department of Education.

Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989). *A world of difference: An international assessment of mathematics and science*. Princeton: Educational Testing Service.

Li, D., & Oranje, A. (2007). *Estimation of standard errors of regression effects in latent regression models using Binder's linearization* (Research Report No. RR-07-09). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02051.x

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large scale assessments. *Journal of Educational and Behavioral Statistics, 34*, 433–463. https://doi.org/10.3102/1076998609332757

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102. https://doi.org/10.1207/s15324818ame0601_5
Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8*, 135–155.
Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley–Interscience. https://doi.org/10.1002/9781119013563
Longley, J. W. (1967). An appraisal of least-squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association, 62*, 819–841. https://doi.org/10.1080/01621459.1967.10500896
Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242. https://doi.org/10.1007/BF02297844
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
Martin, M. O., & Kelly, D. L. (Eds.). (1996). *TIMSS technical report, Volume I: Design and development*. Chestnut Hill: Boston College.
Mayeske, G. W., & Beaton, A. E. (1975). *Special studies of our nation's students*. Washington, DC: U.S. Government Printing Office.
Mayeske, G. W., Cohen, W. M., Wisler, C. E., Okada, T., Beaton, A. E., Proshek, J. M., et al. (1972). *A study of our nation's schools*. Washington, DC: U.S. Government Printing Office.
Mayeske, G. W., Okada, T., & Beaton, A. E. (1973a). *A study of the attitude toward life of our nation's students*. Washington, DC: U.S. Government Printing Office.
Mayeske, G. W., Okada, T., Beaton, A. E., Cohen, W. M., & Wisler, C. E. (1973b). *A study of the achievement of our nation's students*. Washington, DC: U.S. Government Printing Office.
McLaughlin, D. H. (1998). *Study of the linkages of 1996 NAEP and state mathematics assessments in four states*. Washington, DC: National Center for Education Statistics.
McLaughlin, D. H. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates* (Report to the National Institute of Statistical Sciences). Palo Alto: American Institutes for Research.
McLaughlin, D. H. (2005). *Properties of NAEP full population estimates*. Palo Alto: American Institutes for Research.
Messick, S. (1986). *Large-scale educational assessment as policy research: Aspirations and limitations* (Research Report No. RR-86-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1986.tb00182.x
Messick, S., Beaton, A. E., & Lord, F. (1983). *A new design for a new era*. Princeton: Educational Testing Service.
Milton, R. C., & Nelder, J. A. (Eds.). (1969). *Statistical computation*. Waltham: Academic Press.
Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381. https://doi.org/10.1007/BF02306026
Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993–997. https://doi.org/10.1080/01621459.1985.10478215
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196. https://doi.org/10.1007/BF02294457
Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. Princeton: Educational Testing Service.
Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Chicago: Scientific Software.
Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (No. 15–TR–20, pp. 293–360). Princeton: Educational Testing Service.
Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*, 131–154. https://doi.org/10.3102/10769986017002131

Moran, R., & Dresher, A. (2007, April). *Results from NAEP marginal estimation research on multivariate scales.* Paper presented at the meeting of the National Council for Measurement in Education, Chicago, IL.

Mosteller, F., & Moynihan, D. P. (1972). A pathbreaking report: Further studies of the Coleman report. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 3–68). New York: Vintage Books.

Mosteller, F., Fienberg, S. E., Hoaglin, D. C., & Tanur, J. M. (Eds.). (2010). *The pleasures of statistics: The autobiography of Frederick Mosteller*. New York: Springer.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill: International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data* [Computer software]. Chicago: Scientific Software.

Muraki, E., Hombo, C. M., & Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337. https://doi.org/10.1177/01466210022031787

National Assessment of Educational Progress. (1985.) *The reading report card: Progress toward excellence in our school: Trends in reading over four national assessments, 1971-1984* (NAEP Report No. 15-R-01). Princeton: Educational Testing Service.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U. S. Government Printing Office.

Newton, R. G., & Spurrell, D. J. (1967a). A development of multiple regression for the analysis of routine data. *Applied Statistics, 16*, 51–64. https://doi.org/10.2307/2985237

Newton, R. G., & Spurrell, D. J. (1967b). Examples of the use of elements for clarifying regression analyses. *Applied Statistics, 16*, 165–172.

No Child Left Behind Act, P.L. 107-110, 115 Stat. § 1425 (2002).

Oranje, A. (2006a). *Confidence intervals for proportion estimates in complex samples* (Research Report No. RR-06-21). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02027.x

Oranje, A. (2006b). *Jackknife estimation of sampling variance of ratio estimators in complex samples: Bias and the coefficient of variation* (Research Report No. RR-06-19). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02025.x

Oranje, A., Freund, D., Lin, M.-J., & Tang, Y. (2007). *Disclosure risk in educational surveys: An application to the National Assessment of Educational Progress* (Research Report No. RR-07-24). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02066.x

Oranje, A., Li, D., & Kandathil, M. (2009). *Evaluation of methods to compute complex sample standard errors in latent regression models* (Research Report No. RR-09-49). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02206.x

Organisation for Economic Co-operation and Development. (2013). *OECD skills outlook 2013: First results from the survey of adult skills*. Paris: OECD Publishing.

Pashley, P. J., & Phillips, G. W. (1993). *Toward world-class standards: A research study linking international and national assessments*. Princeton: Educational Testing Service.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando: Harcourt Brace.

Phillips, G. (2007). *Chance favors the prepared mind: Mathematics and science indicators for comparing states and nations*. Washington, DC: American Institutes for Research.

Privacy Act, 5 U.S.C. § 552a (1974).

Qian, J. (1998). Estimation of the effective degree of freedom in t-type tests for complex data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704–708. Retrieved from http://www.amstat.org/sections/srms/Proceedings/

Qian, J., & Kaplan, B. (2001). Analysis of design effects for NAEP combined samples. *2001 Proceedings of the American Statistical Association, Survey Research Methods Section* [CD–ROM]. Alexandria: American Statistical Association.

Qian, J., Kaplan, B., & Weng, V. (2003) *Analysis of NAEP combined national and state samples* (Research Report No. RR-03-21). Princeton: Educational Testing Service.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43*, 353–360. https://doi.org/10.1093/biomet/43.3-4.353

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage.

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large–Scale Assessment, 4*, 59–74.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rock, D. A., Hilton, T., Pollack, J. M., Ekstrom, R., & Goertz, M. E. (1985). *Psychometric analysis of the NLS-72 and the high school and beyond test batteries* (NCES Report No. 85-218). Washington, DC: National Center for Education Statistics.

Rogers, A., Tang, C., Lin, M. J., & Kandathil, M. (2006). DGROUP [Computer software]. Princeton: Educational Testing Service.

Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425–435. https://doi.org/10.1007/BF02306030

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72*, 538–543. https://doi.org/10.1080/01621459.1977.10480610

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley. https://doi.org/10.1002/9780470316696

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: CRC Press.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6*, 309–316. https://doi.org/10.1007/BF02288586

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions* (Research Report No. RR-05-27). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02004.x

Sinharay, S., Guo, Z., von Davier, M., & Veldkamp, B. P. (2010). Assessing fit of latent regression models. *IERI Monograph Series, 3*, 35–55.

Statistics Canada & Organisation for Economic Co-operation and Development. (2005). *Learning a living: First results of the adult literacy and life skills survey*. Paris: OECD Publishing.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–312). New York: Springer. https://doi.org/10.1007/978-0-387-49771-6_16

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309–322. https://doi.org/10.2307/1390648

Thomas, N. (2000). Assessing model sensitivity of imputation methods used in NAEP. *Journal of Educational and Behavioral Statistics, 25*, 351–371. https://doi.org/10.3102/10769986025004351

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*, 33–48. https://doi.org/10.1007/BF02294708

Thorndike, R. L., & Hagen, E. (1959). *Ten thousand careers*. New York: Wiley.

Tukey, J. W. (1958). Bias and confidence in not–quite large samples [abstract]. *The Annals of Mathematical Statistics, 29*, 614.

Viadero, D. (2006). Fresh look at Coleman data yields different conclusions. *Education Week, 25*(41), 21.

von Davier, M. (2003). *Comparing conditional and marginal direct estimation of subgroup distributions* (Research Report No. RR-03-02). Princeton: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01894.x

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32*, 233–251. https://doi.org/10.3102/1076998607300422

von Davier, M., & Sinharay, S. (2010). Stochastic approximation for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*, 174–193. https://doi.org/10.3102/1076998609346970

von Davier, M., & Yon, H. (2004, April) *A conditioning model with relaxed assumptions.* Paper presented at the meeting of the National Council of Measurement in Education, San Diego, CA.

von Davier, M., & Yu, H. T. (2003, April). *Recovery of population characteristics from sparse matrix samples of simulated item responses.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. E. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1056). Amsterdam: Elsevier.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series, 2*, 9–36.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1–21. https://doi.org/10.1111/j.1745-3984.1993.tb00419.x

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109–128. https://doi.org/10.1177/0146621602026001007

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M.A., & Lord, F. M. (1982). LOGIST user's guide Logist 5, version 1.0 [Computer software manual]. Princeton: Educational Testing Service.

Wirtz, W. (Ed.). (1977). *On further examination: Report of the advisory panel on the scholastic aptitude test score decline* (Report No. 1977-07-01). New York: College Entrance Examination Board.

Wirtz, W., & Lapointe, A. (1982). Measuring the quality of education: A report on assessing educational progress. *Educational Measurement: Issues and Practice*, *1*, 17–19, 23. https://doi.org/10.1111/j.1745-3992.1982.tb00673.x

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST – A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum No. RM-76-06). Princeton: Educational Testing Service.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*, 155–173. https://doi.org/10.2307/1165167

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association, 57*, 348–368. https://doi.org/10.1080/01621459.1962.10480664

Zwick, R. (1987a). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*, 293–308. https://doi.org/10.1111/j.1745-3984.1987.tb00281.x

Zwick, R. (1987b). Some properties of the correlation matrix of dichotomous Guttman items. *Psychometrika, 52*, 515–520. https://doi.org/10.1007/BF02294816

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10–16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x