Part II
ETS Contributions to Education Policy and Evaluation

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,…, $2^N N!$,), $y_k = X\beta + e_{\mathrm{k}},$ 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

Chapter 9
Large-Scale Assessments of Adult Literacy

Irwin Kirsch, Mary Louise Lennon, Kentaro Yamamoto, and Matthias von Davier

Educational Testing Service's (ETS's) work in large-scale adult literacy assessments has been an ongoing and evolving effort, beginning in 1984 with the Young Adult Literacy Survey in the United States. This work has been designed to meet policy needs, both in the United States and internationally, based on the growing awareness of literacy as human capital. The impact of these assessments has grown as policy makers and other stakeholders have increasingly come to understand the critical role that foundational skills play in allowing individuals to maintain and enhance their ability to meet changing work conditions and societal demands. For example, findings from these surveys have provided a wealth of information about how the distribution of skills is related to social and economic outcomes. Of equal importance, the surveys and associated research activities have contributed to large-scale assessment methodology, the development of innovative item types and delivery systems, and methods for reporting survey data in ways that ensure its utility to a range of stakeholders and audiences.

The chronology of ETS's large-scale literacy assessments, as shown in Fig. 9.1, spans more than 30 years. ETS served as the lead contractor in the development of these innovative assessments, while the prime clients and users of the assessment outcomes were representatives of either governmental organizations such as the National Center for Education Statistics (NCES) and Statistics Canada, or transgovernmental entities such as the Organisation for Economic Co-operation and Development (OECD). These instruments have evolved from a single-language, paper-based assessment focusing on a U.S. population of 16- to 25-year-olds to an adaptive, computer-based assessment administered in almost 40 countries and close to 50 languages to adults through the age of 65. By design, the assessments have been linked at the item level, with sets of questions from previous assessments

This work was conducted while M. von Davier was employed with Educational Testing Service.

I. Kirsch (✉) • M.L. Lennon • K. Yamamoto • M. von Davier
Educational Testing Service, Princeton, NJ, USA
e-mail: ikirsch@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_9

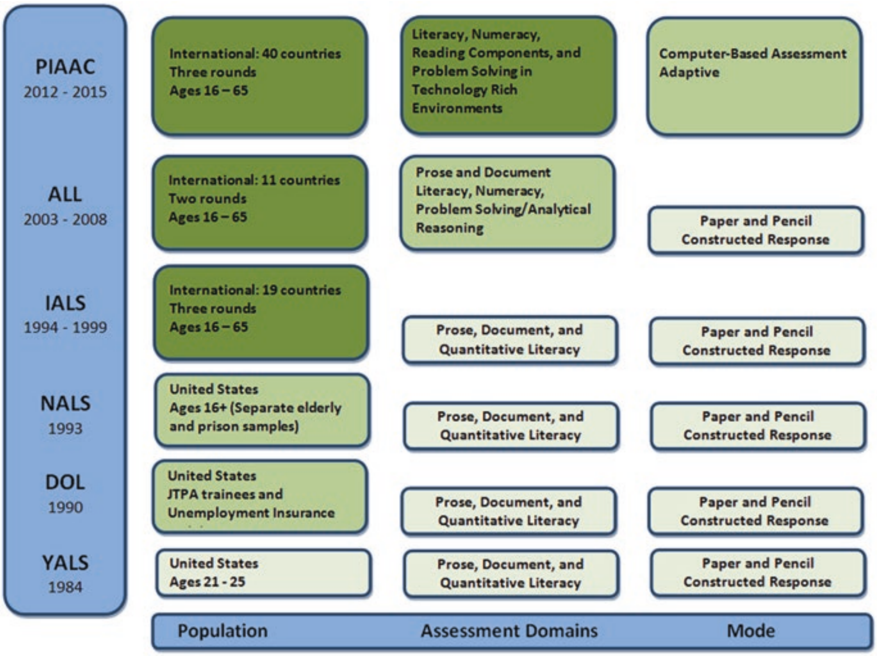

PIAAC
2012 - 2015
International: 40 countries
Three rounds
Ages 16 – 65
Literacy, Numeracy, Reading Components, and Problem Solving in Technology Rich Environments
Computer-Based Assessment
Adaptive
ALL
2003 - 2008
International: 11 countries
Two rounds
Ages 16 – 65
Prose and Document Literacy, Numeracy, Problem Solving/Analytical Reasoning
Paper and Pencil
Constructed Response
IALS
1994 - 1999
International: 19 countries
Three rounds
Ages 16 – 65
Prose, Document, and Quantitative Literacy
Paper and Pencil
Constructed Response
NALS
1993
United States
Ages 16+ (Separate elderly and prison samples)
Prose, Document, and Quantitative Literacy
Paper and Pencil
Constructed Response
DOL
1990
United States
JTPA trainees and Unemployment Insurance
Prose, Document, and Quantitative Literacy
Paper and Pencil
Constructed Response
YALS
1984
United States
Ages 21 - 25
Prose, Document, and Quantitative Literacy
Paper and Pencil
Constructed Response
Population
Assessment Domains
Mode

Fig. 9.1 ETS's large-scale literacy assessments. Note. *ALL* = Adult Literacy and Life Skills Survey (Statistics Canada, Organisation for Economic Co-operation and Development [OECD]), *DOL* = Department of Labor Survey, *JPTA* = Job Training Partnership Act, *IALS* = International Adult Literacy Survey (Statistics Canada, OECD), *PIAAC* = Programme for the International Assessment of Adult Competencies (OECD), *YALS* = Young Adult Literacy Survey (through the National Assessment of Educational Progress)

included in each new survey. This link has made it possible to look at changes in skill levels, as well as the distribution of those skills, over time. Each of the assessments has also expanded upon previous surveys. As Fig. 9.1 illustrates, the assessments have changed over the years in terms of who is assessed, what skills are assessed, and how those skills are assessed. The surveys have evolved to include larger and more diverse populations as well as new and expanded constructs. They have also evolved from a paper-and-pencil, open-ended response mode to an adaptive, computer-based assessment.

In many ways, as the latest survey in this 30-year history, the Programme for the International Assessment of Adult Competencies (PIAAC) represents the culmination of all that has been learned over several decades in terms of instrument design, translation and adaptation procedures, scoring, and the development of interpretive schemes. As the first computer-based assessment to be used in a large-scale household skills survey, the experience derived from developing and delivering PIAAC—including research focused on innovative item types, harvesting log files, and delivering an adaptive assessment—helped lay the foundation for new computer based large-scale assessments yet to come.

This paper describes the contributions of ETS to the evolution of large-scale adult literacy assessments in six key areas:

- Expanding the construct of literacy
- Developing a model for building construct-based assessments
- Expanding and implementing large-scale assessment methodology
- Linking real-life stimulus materials and innovative item types
- Developing extensive background questionnaires to link performance with experience and outcome variables
- Establishing innovative reporting procedures to better integrate research and survey data

9.1 Expanding the Construct of Literacy

Early work in the field of adult literacy defined literacy based on the attainment of certain grade level scores on standardized academic tests of reading achievement. Standards for proficiency increased over the decades with "functional literacy" being defined as performance at a fourth-grade reading level during World War II, eighth-grade level in the 1960s, and a 12th grade level by the early 1970s. This grade-level focus using instruments that consisted of school-based materials was followed by a competency-based approach that employed tests based on nonschool materials from adult contexts. Despite this improvement, these tests still viewed literacy along a single continuum, defining individuals as either *literate* or *functionally illiterate* based on where they performed along that continuum. The 1984 Young Adult Literacy Survey (YALS) was the first in a series of assessments that contributed to an increasingly broader understanding of what it means to be "literate" in complex modern societies. In YALS, the conceptualization of literacy was expanded to reflect the diversity of tasks that adults encounter at work, home, and school and in their communities. As has been the case for all of the large-scale literacy assessments, panels of experts were convened to help set the framework for this assessment. Their deliberations led to the adoption of the following definition of literacy: "using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential" (Kirsch and Jungeblut 1986, p. 3).

This definition both rejected an arbitrary standard for literacy, such as performing at a particular grade level on a test of reading, and implied that literacy comprises a set of complex information-processing skills that goes beyond decoding and comprehending text-based materials.

To better reflect this multi-faceted set of skills and abilities, performance in YALS was reported across three domains, defined as follows (Kirsch and Jungeblut 1986, p. 4):

- Prose literacy: the knowledge and skills needed to understand and use information from texts including editorials, news stories, poems, and the like
- Document literacy: the knowledge and skills required to locate and use information contained in job applications or payroll forms, bus schedules, maps, indexes, and so forth

- Quantitative literacy: the knowledge and skills required to apply arithmetic operations, either alone or sequentially, that are embedded in printed materials, such as in balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest on a loan from an advertisement

Rather than attempt to categorize individuals, or groups of individuals, as literate or illiterate, YALS reported results for each of these three domains by characterizing the underlying information-processing skills required to complete tasks at various points along a 0–500-point reporting scale, with a mean of 305 and a standard deviation of about 50. This proficiency-based approach to reporting was seen as a more faithful representation of both the complex nature of literacy demands in society and the various types and levels of literacy demonstrated by young adults.

Subsequent research at ETS led to the definition of five levels within the 500-point scale. Analyses of the interaction between assessment materials and the tasks based on those materials defined points along the scale at which information-processing demands shifted. The resulting levels more clearly delineated the progression of skills required to complete tasks at different points on the literacy scales and helped characterize the skills and strategies underlying the prose, document, and quantitative literacy constructs. These five levels have been used to report results for all subsequent literacy surveys, and the results from each of those assessments have made it possible to further refine our understanding of the information-processing demands at each level as well as the characteristics of individuals performing along each level of the scale.[1]

With the 2003 Adult Literacy and Life Skills Survey (ALL), the quantitative literacy domain was broadened to reflect the evolving perspective of experts in the field. The new numeracy domain was defined as the ability to interpret, apply, and communicate numerical information. While quantitative literacy focused on quantitative information embedded in text and primarily required respondents to demonstrate computational skills, numeracy included a broader range of skills typical of many everyday and work tasks including sorting, measuring, estimating, conjecturing, and using models. This expanded domain allowed ALL to collect more information about how adults apply mathematical knowledge and skills to real-life situations. In addition, the ALL assessment included a problem-solving component that focused on analytical reasoning. This component collected information about the ability of adults to solve problems by clarifying the nature of a problem and developing and applying appropriate solution strategies. The inclusion of problem solving was seen as a way to improve measurement at the upper levels of the scales and to reflect a skill set of growing interest for adult populations.

Most recently, the concept of literacy was expanded again with the Programme for the International Assessment of Adult Competencies (PIAAC). As the first computer-based, large-scale adult literacy assessment, PIAAC reflected the changing nature of information, its role in society, and its impact on people's lives.

[1] See the appendix for a description of the information-processing demands associated with each of the five levels across the literacy domains.

The scope of the prose, document, and numeracy domains was broadened in PIAAC and the assessment incorporated two new domains, as follows:

- For the first time, this adult assessment addressed literacy in digital environments. As a computer-based assessment, PIAAC included tasks that required respondents to use electronic texts including web pages, e-mails, and discussion boards. These stimulus materials included hypertext and multiple screens of information and simulated real-life literacy demands presented by digital media.
- In PIAAC, the definition of numeracy was broadened again to include the ability to access, use, interpret, and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life. The inclusion of *engage* in the definition signaled that not only cognitive skills but also dispositional elements (i.e., beliefs and attitudes) are necessary to meet the demands of numeracy effectively in everyday life.
- PIAAC included the new domain of problem-solving in technology-rich environments (PS-TRE), the first attempt to assess this domain on a large scale and as a single dimension. PS-TRE was defined as:

  > using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks. The first PIAAC problem-solving survey focuses on the abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks. (OECD 2012, p. 47)

  PS-TRE presented computer-based tasks designed to measure the ability to analyze various requirements of a task, define goals and plans, and monitor progress until the task purposes were achieved. Simulated web, e-mail and spreadsheet environments were created and respondents were required to use multiple, complex sources of information, in some cases across more than one environment, to complete the presented tasks. The focus of these tasks was not on computer skills per se, but rather on the cognitive skills required to access and make use of computer-based information to solve problems.
- Finally, PIAAC contained a reading components domain, which included measures of vocabulary knowledge, sentence processing, and passage comprehension. Adding this domain was an important evolution because it provided more information about the skills of individuals with low levels of literacy proficiency than had been available from previous international assessments. To have a full picture of literacy in any society, it is necessary to have more information about these individuals because they are at the greatest risk of negative social, economic, and labor market outcomes.

9.2 Developing a Model for Building Construct-Based Assessments

A key characteristic of the large-scale literacy assessments is that each was based on a framework that, following Messick's (1994) construct-centered approach, defined the construct to be measured, the performances or behaviors expected to reveal that construct, and the characteristics of assessment tasks to elicit those behaviors. In the course of developing these assessments, a model for the framework development process was created, tested, and refined. This six-part process, as shown in Fig. 9.2 and described in more detail below, provides a logical sequence of steps from clearly defining a particular skill area to developing specifications for item construction and providing a foundation for an empirically based interpretation of the assessment results. Through this process, the inferences and assumptions about what is to be measured and how the results will be interpreted and reported are explicitly described.

1. *Develop a general definition of the domain.* The first step in this model is to develop a working definition of the domain and the assumptions underlying it. It is this definition that sets the boundaries for what will and will not be measured in a given assessment.
2. *Organize the domain.* Once the definition is developed, it is important to think about the kinds of tasks that represent the skills and abilities included in that

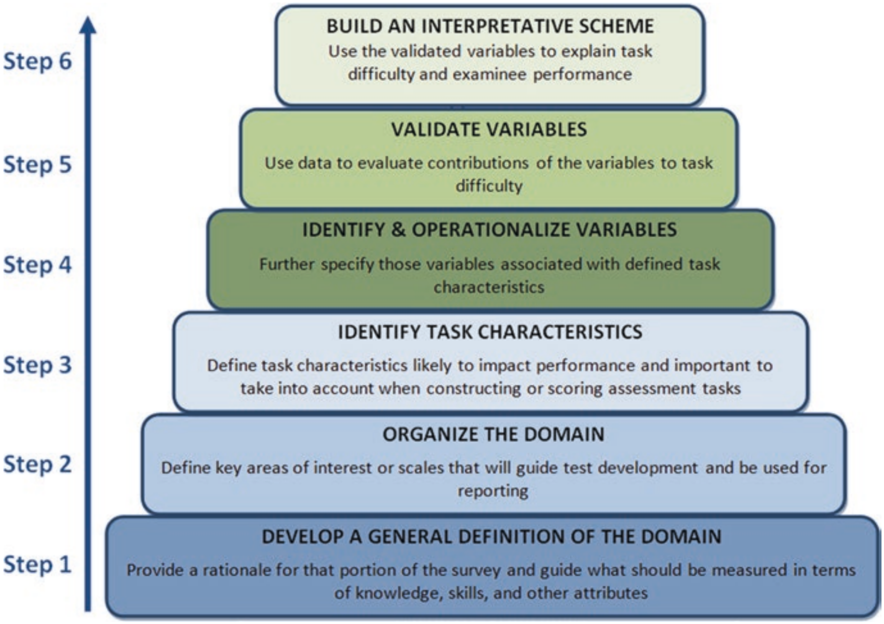

Step 6
BUILD AN INTERPRETATIVE SCHEME
Use the validated variables to explain task difficulty and examinee performance
Step 5
VALIDATE VARIABLES
Use data to evaluate contributions of the variables to task difficulty
Step 4
IDENTIFY & OPERATIONALIZE VARIABLES
Further specify those variables associated with defined task characteristics
Step 3
IDENTIFY TASK CHARACTERISTICS
Define task characteristics likely to impact performance and important to take into account when constructing or scoring assessment tasks
Step 2
ORGANIZE THE DOMAIN
Define key areas of interest or scales that will guide test development and be used for reporting
Step 1
DEVELOP A GENERAL DEFINITION OF THE DOMAIN
Provide a rationale for that portion of the survey and guide what should be measured in terms of knowledge, skills, and other attributes

Fig. 9.2 Model for construct-based assessment

definition. Those tasks must then be categorized in relation to the construct definition to inform test design and result in meaningful score reporting. This step makes it possible to move beyond a laundry list of tasks or skills to a coherent representation of the domain that will permit policy makers and others to summarize and report information in more useful ways.

3. *Identify task characteristics*. Step 3 involves identifying a set of key characteristics, or task models, which will be used in constructing tasks for the assessment. These models may define characteristics of the stimulus materials to be used as well as characteristics of the tasks presented to examinees. Examples of key task characteristics that have been employed throughout the adult literacy assessments include contexts, material types, and information-processing demands.
4. *Identify and operationalize variables.* In order to use the task characteristics in designing the assessment and, later, in interpreting the results, the variables associated with each task characteristic need to be defined. These definitions are based on the existing literature and on experience with building and conducting other large-scale assessments. Defining the variables allows item developers to categorize the materials with which they are working, as well as the questions and directives they construct, so that these categories can be used in the reporting of the results. In the literacy assessments, for example, *context* has been defined to include home and family, health and safety, community and citizenship, consumer economics, work, leisure, and recreation; *materials* have been divided into continuous and noncontinuous texts with each of those categories being further specified; and *processes* have been identified in terms of type of match (focusing on the match between a question and text and including locating, integrating and generating strategies), type of information requested (ranging from concrete to abstract), and plausibility of distractors.[2]
5. *Validate variables*. In Step 5, research is conducted to validate the variables used to develop the assessment tasks. Statistical analyses determine which of the variables account for large percentages of the variance in the difficulty distribution of tasks and thereby contribute most towards understanding task difficulty and predicting performance. In the literacy assessments, this step provides empirical evidence that a set of underlying process variables represents the skills and strategies involved in accomplishing various kinds of literacy tasks.
6. *Build an interpretative scheme.* Finally in Step 6, an interpretative scheme is built that uses the validated variables to explain task difficulty and examinee performance. The definition of proficiency levels to explain performance along the literacy scales is an example of such an interpretative scheme. As previously explained, each scale in the literacy assessments has been divided into five progressive levels characterized by tasks of increasing complexity, as defined by the underlying information processing demands of the tasks. This scheme has been used to define what scores along a particular scale mean and to describe the survey results. Thus, it contributes to the construct validity of inferences based

[2] See Kirsch (2001) and Murray et al. (1997) for a more detailed description of the variables used in the IALS and subsequent assessments.

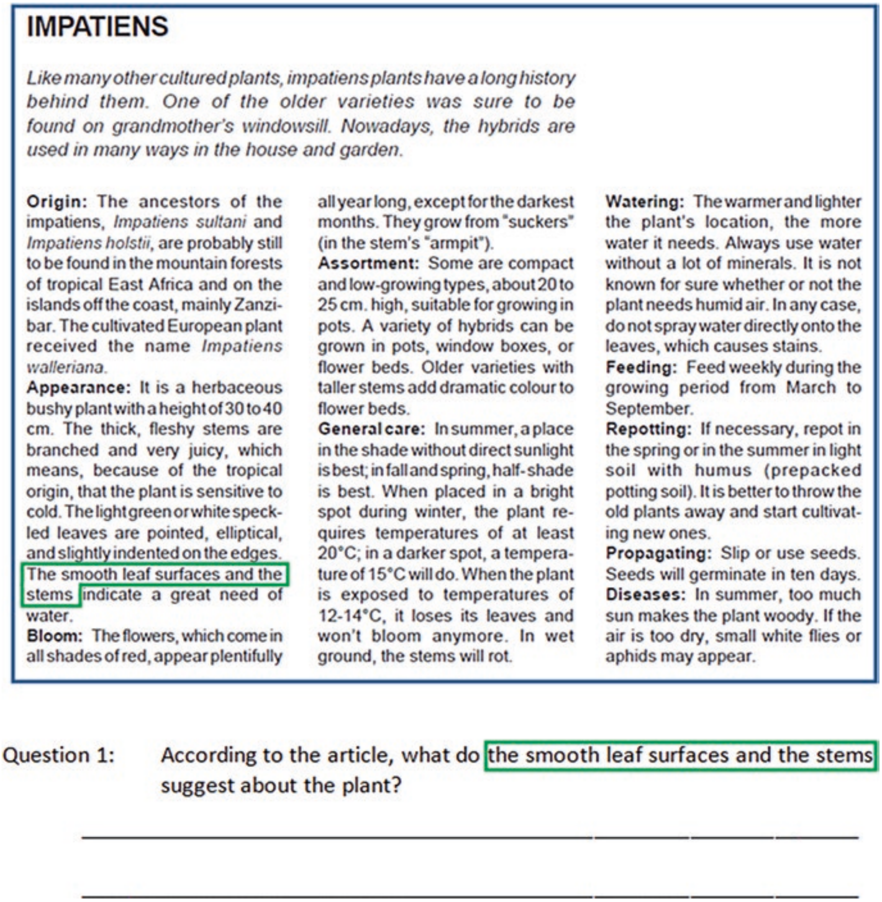
on scores from the measure (Messick 1989). Data from the surveys' background questionnaires have demonstrated consistent correlations between the literacy levels and social and economic outcomes, providing additional evidence for the validity of this particular scheme.

Advancing Messick's approach to construct-based assessment through the application of this framework development model has been one important contribution of the large-scale literacy surveys. This approach not only was used for each of these literacy assessments, but also has become an accepted practice in other assessment programs including the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Achievement (PISA) and the United Nations Educational, Scientific, and Cultural Organization's (UNESCO's) Literacy Assessment and Monitoring Programme (LAMP).

Employing this model across the literacy assessments both informed the test development process and allowed ETS researchers to explore variables that explained differences in performance. Research based on data from the early adult literacy assessments led to an understanding of the relationship between the print materials that adults use in their everyday lives and the kinds of tasks they need to accomplish using such materials. Prior difficulty models for both assessments and learning materials tended to focus on the complexity of stimulus materials alone. ETS's research focused on both the linguistic features and the structures of prose and document materials, as well as a range of variables related to task demands.

Analyses of the linguistic features of stimulus materials first identified the important distinction between continuous and noncontinuous texts. Continuous texts (the prose materials used in the assessments) are composed of sentences that are typically organized into paragraphs. Noncontinuous texts (document materials) are more frequently organized in a matrix format, based on combinations of lists. Work by Mosenthal and Kirsch (1991) further identified a taxonomy of document structures that organized the vast range of matrix materials found in everyday life—television schedules, checkbook registers, restaurant menus, tables of interest rates, and so forth—into six structures: simple, combined, intersecting, and nested lists; and charts and graphs. In prose materials, analyses of the literacy data identified the impact of features such as the presence or absence of graphic organizers including headings, bullets, and bold or italicized print.

On the task side of the difficulty equation, these analyses also identified strategies required to match information in a question or directive with corresponding information in prose and document materials. These strategies—locate, cycle, integrate, and generate—in combination with text features, helped explain what made some tasks more or less difficult than others (Kirsch 2001). For example, locate tasks were defined as those that required respondents to match one or more features of information stated in the question to either identical or synonymous information in the stimulus. A locate task could be fairly simple if there was an exact match between the requested information in the question or directive and the wording in the stimulus and if the stimulus was relatively short, making the match easy to find.



IMPATIENS

Like many other cultured plants, impatiens plants have a long history behind them. One of the older varieties was sure to be found on grandmother's windowsill. Nowadays, the hybrids are used in many ways in the house and garden.

Origin: The ancestors of the impatiens, *Impatiens sultani* and *Impatiens holstii*, are probably still to be found in the mountain forests of tropical East Africa and on the islands off the coast, mainly Zanzibar. The cultivated European plant received the name *Impatiens walleriana*.

Appearance: It is a herbaceous bushy plant with a height of 30 to 40 cm. The thick, fleshy stems are branched and very juicy, which means, because of the tropical origin, that the plant is sensitive to cold. The light green or white speckled leaves are pointed, elliptical, and slightly indented on the edges. The smooth leaf surfaces and the stems indicate a great need of water.

Bloom: The flowers, which come in all shades of red, appear plentifully all year long, except for the darkest months. They grow from "suckers" (in the stem's "armpit").

Assortment: Some are compact and low-growing types, about 20 to 25 cm. high, suitable for growing in pots. A variety of hybrids can be grown in pots, window boxes, or flower beds. Older varieties with taller stems add dramatic colour to flower beds.

General care: In summer, a place in the shade without direct sunlight is best; in fall and spring, half-shade is best. When placed in a bright spot during winter, the plant requires temperatures of at least 20°C; in a darker spot, a temperature of 15°C will do. When the plant is exposed to temperatures of 12-14°C, it loses its leaves and won't bloom anymore. In wet ground, the stems will rot.

Watering: The warmer and lighter the plant's location, the more water it needs. Always use water without a lot of minerals. It is not known for sure whether or not the plant needs humid air. In any case, do not spray water directly onto the leaves, which causes stains.

Feeding: Feed weekly during the growing period from March to September.

Repotting: If necessary, repot in the spring or in the summer in light soil with humus (prepacked potting soil). It is better to throw the old plants away and start cultivating new ones.

Propagating: Slip or use seeds. Seeds will germinate in ten days.

Diseases: In summer, too much sun makes the plant woody. If the air is too dry, small white flies or aphids may appear.

Question 1: According to the article, what do the smooth leaf surfaces and the stems suggest about the plant?

____

____

Fig. 9.3 Sample prose task

As an example, see Fig. 9.3. Here there is an exact match between "the smooth leaf surfaces and the stems" in the question and in the last sentence in the second paragraph of the text.

Analyses showed that the difficulty of locate tasks increased when stimuli were longer and more complex, making the requested information more difficult to locate; or when there were distractors, or a number of plausible correct answers, within the text. Difficulty also increased when requested information did not exactly match the text in the stimulus, requiring respondents to locate synonymous information. By studying and defining the interaction between the task demands for locate, cycle, integrate, and generate tasks and features of various stimuli, the underlying information-processing skills could be more clearly understood. This research allowed for improved assessment design, increased interpretability of results, and development of derivative materials, including individual assessments[3] and instructional materials.[4]

In 1994, the literacy assessments moved from a national to an international focus. The primary goal of the international literacy assessments—International Adult Literacy Survey (IALS), ALL, and PIAAC—was to collect comparable international data that would provide a broader understanding of literacy across industrialized nations.

One challenge in meeting the goal of ensuring comparability across different national versions of the assessment was managing the translation process. Based on the construct knowledge gained from earlier assessments, it was clear that translators had to understand critical features of both the stimulus materials and the questions. Training materials and procedures were developed to help translators and project managers from participating countries reach this understanding. For example, the translation guidelines for the content shown in Fig. 9.3 specified the following:

- Translation must maintain literal match between the key phrase "the smooth leaf surfaces and the stems" in the question and in the last sentence in the second paragraph of the text.
- Translation must maintain a synonymous match between *suggest* in question and *indicate* in text.

Understanding task characteristics and the interaction between questions and stimulus materials allowed test developers to create precise translation guidelines to ensure that participating countries developed comparable versions of the assessment instruments. The success of these large-scale international efforts was in large part possible because of the construct knowledge gained from ETS research based on the results of earlier national assessments.

9.3 Expanding and Implementing Large-Scale Assessment Methodology

The primary purpose of the adult literacy large-scale assessments has been to describe the distribution of literacy skills in populations, as well as in subgroups within and across populations. The assessments have not targeted the production of

[3] These individual assessments include the Test of Applied Literacy Skills (TALS), a paper-and-pencil assessment with multiple forms; the *PDQ Profile*™ Series, an adaptive computer-based assessment of literacy proficiency; and the Health Activities Literacy Test, an adaptive computer-based assessment of literacy tasks focusing on health issues.

[4] Using information from this research, ETS developed P.D.Q. Building Skills for Using Print in the early 1990s. This multi-media, group-based system includes more than 100 h of instruction focusing on prose, document, and quantitative literacy, as well as workbooks and instructional support materials.

scores for individual test takers, but rather employed a set of specialized design principles and statistical tools that allow a reliable and valid description of skill distributions for policy makers and other stakeholders. To describe skills in a comparable manner in international contexts, the methodologies utilized needed to ensure that distributions were reported in terms of quantities that describe differences on scales across subgroups in meaningful ways for all participating entities.

The requirement to provide comparable estimates of skill distributions has been met by using the following methodological tools:

- Models that allow the derivation of comparable measures across populations and comparisons across literacy assessments
- Survey methodologies that provide representative samples of respondents
- Procedures to ensure scoring accuracy and to handle missing data
- Forward-looking designs that take advantage of context information in computer-based assessments

Taken together, these methodological tools facilitate the measurement goal of providing reliable, valid, and comparable estimates of skill distributions based on large-scale literacy assessments.

9.3.1 Models Allowing the Derivation of Comparable Measures and Comparisons Across Literacy Assessments

The goal of the literacy assessments discussed here has been to provide a description of skills across a broad range of ability, particularly given that the assessments target adults who have very different educational backgrounds and a wider range of life experiences than school-based populations. Thus the assessments have needed to include tasks that range from very easy to very challenging. To enable comparisons across a broad range of skill levels and tasks, the designs for all of the adult literacy assessments have used “incomplete block designs”. In such designs, each sampled individual takes a subset of the complete assessment. The method of choice for the derivation of comparable measures in incomplete block designs is based on measurement models that were developed for providing such measures in the analyses of test data (Lord 1980; Rasch 1960). These measurement models are now typically referred to as item response theory (IRT) models (Lord and Novick 1968).

IRT models are generally considered superior to simpler approaches based on sum scores, particularly in the way omitted responses and incomplete designs can be handled. Because IRT uses the full information contained in the set of responses, these models are particularly useful for assessment designs that utilize a variety of item types arranged in blocks that cannot be set up to be parallel forms of a test. Incomplete block designs do not allow the comparison of sum scores of aggregated responses because different blocks of items may vary in difficulty and even in the number of items. IRT models establish a comparable scale on which items from different blocks, and from respondents taking different sets of items, can be located, even in sparse incomplete designs. These models are powerful tools to evaluate whether the information provided for each individual item is comparable across populations of interest (see, for example, Yamamoto and Mazzeo 1992). In particular, the linking procedures typically used in IRT have been adapted, refined, and generalized for use in international assessments of adult literacy. More specifically, recent developments in IRT linking methods allow a more flexible approach to the alignment of scales that takes into account local deviations (Glas and Verhelst 1995; Yamamoto 1998; von Davier and von Davier 2007; Oliveri and von Davier 2011; Mazzeo and von Davier 2014; Glas and Jehangir 2014). The approach applied in IALS, ALL and PIAAC enables international assessments to be linked across a large number of common items while allowing for a small subset of items in each country to function somewhat differently to eliminate bias due to occasional item-by-country interactions. IRT has been the measurement method of choice not only for ETS's adult literacy assessments, but also for national and international assessments of school-age students such as the National Assessment of Educational Progress (NAEP), PISA, and Trends in International Mathematics and Science Study (TIMSS).

The integration of background information is a second important characteristic of the analytical methodologies used in the adult literacy assessments. Background data are used for at least two purposes in this context. First and foremost, they provide information about the relationship between demographic variables and skills. This makes it possible to investigate how the distribution of skills is associated with variables including educational attainment, gender, occupation, and immigration status of groups. These are among the variables needed to answer questions that are of interest to policy makers and other stakeholders, such as, "How are skills distributed in immigrant vs. nonimmigrant populations?" and "What is the relationship between literacy skills and measures of civic engagement such as voting?" In addition, background data provide auxiliary information that can be used to improve the precision of the skills measurement. This use of background data is particularly important because the available background data can help alleviate the effects of limited testing time for respondents by using the systematic differences between groups of respondents to strengthen the estimation of skills.[5]

While one of the main aims of ETS's large-scale literacy assessments has been to provide data on human capital at any given point in time, the extent to which skills change over time is also of fundamental interest. IRT models provide a powerful tool to link assessments over cycles conducted in different years. In much the same way that IRT allows linking of scales and provides comparable measures across blocks of different items within an assessment, and across countries, IRT can also be used to link different assessments over time. This link is only possible because significant efforts have been made across the literacy assessments to collect data in a manner that supports reusing sets of items over time while regularly renew-

[5] The interested reader is referred to Mislevy et al. (1992) for a description of this approach and to von Davier et al. (2006) for an overview and a description of recent improvements and extensions of the approach.

ing the item pool. The particular design principles applied ensure that new and previously used blocks of items are combined into test booklets in such a way that each assessment is also connected to multiple assessments over time. Because IRT estimation methods have been developed and extended to facilitate analyses of incomplete designs, these methods are particularly well suited to analyze multiple links across assessments. Statistical tools can be used to evaluate whether the items used repeatedly in multiple assessments are indeed comparable across assessments from different years and provide guidance as to which items to retain and which parts of the assessment have to be renewed by adding new task material.

9.3.2 Survey Methodologies That Provide Representative Samples of Respondents

The description of populations with respect to policy-relevant variables requires that members of the population of interest are observed with some positive probability. While it is not a requirement (or possibility) to assess every individual, a representative sample has to be drawn in order to provide descriptions of populations without bias. The adult literacy assessments have typically used methods common to household surveys, in which either a central registry of inhabitants or a list of addresses of dwellings/households of a country is used to randomly draw a representative random sample of respondents. This list is then used to select an individual at random, get in contact with those selected and ask the selected individual to participate in the survey. To account for unequal chances of being selected, the use of sampling weights is necessary. The importance of sampling and weighting for an accurate estimate of skill distributions is discussed in more detail in contributions summarizing analytic strategies involving sampling and weights for large-scale assessments by Rust (2014) and Rutkowski et al. (2010).

One particular use of these survey methodologies in large-scale assessments, and a contribution of ETS's adult assessments, is the projection of skill distributions based on expected changes in the population. The report, *America's Perfect Storm: Three Forces Changing Our Nation's Future* (Kirsch et al. 2007) shows how evidence regarding skill distributions in populations of interest can be projected to reflect changes in those populations, allowing a prediction of the increase or decline of human capital over time.

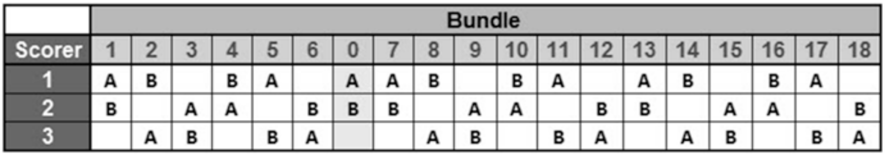
9.3.3 Procedures to Ensure Scoring Accuracy

One measurement issue that has been addressed in large-scale literacy assessments is the need to ensure that paper-and-pencil (as well as human-scored computer-based) tasks are scored accurately and reliably, both within and across countries participating in the international surveys. Many of the assessment tasks require respondents to provide short, written responses that typically range in length from single-word responses to short phrases or sentences. Some tasks ask for responses to be marked on the stimulus. On paper, respondents may be asked to circle or underline the correct answer whereas on the computer, respondents may be required to mark or highlight the response using the mouse or another input device. So while responses are typically quite short, scorers in all participating countries must follow a well-developed set of scoring rules to ensure consistent scoring. All of the adult literacy surveys prior to PIAAC were conducted as paper-and-pencil assessments, scored by national teams of trained scorers. While PIAAC is largely a computer-based assessment using automated scoring, a paper-and-pencil component has been retained, both to strengthen the link between modes and to provide an option for respondents without the requisite technical skills to complete the assessment on the computer. To ensure reliable and comparable data in all of the adult literacy surveys, it was critical that processes were developed to monitor the accuracy of human scoring for the short constructed responses in that mode within a country, across countries, and across assessments over time.

Without accurate, consistent and internationally comparable scoring of paper-and-pencil items, all subsequent psychometric analyses of these items would be severely jeopardized. For all of the large-scale adult literacy assessments, the essential activities associated with maintaining scoring consistency have been basically the same. Having items scored independently by two different scorers and then comparing the resulting scores has been the key required procedure for all participating countries. However, because the number of countries and number of languages has increased with each international assessment, the process has been refined over time. In IALS, the procedure used to ensure standardized scoring involved an exchange of booklets across countries with the same or similar languages. Country A and Country B thus would score their own booklets; then Country A would second score Country B’s booklets and vice versa. In cases where a country could not be paired with another testing in the same language, the scorers within one country would be split into two independent groups, and booklets would be exchanged across groups for rescoring.

Beginning with ALL, the use of anchor booklets was introduced. This common set of booklets was prepared by test developers and distributed to all countries. Item responses in these booklets were based on actual responses collected in the field as well as responses that reflected key points on which scorers were trained. Because responses were provided in English, scoring teams in each country designated two bilingual scorers responsible for the double-scoring process. Anchor booklets were used in PIAAC as well. The new aspect introduced in PIAAC was the requirement that countries follow a specified design to ensure that each booklet was scored twice and that scorers functioned both as first and second scorer across all of the booklets. Figure 9.4 shows the PIAAC design for countries that employed three scorers. The completed booklets were divided up into 18 bundles of equal size. Bundle 0 was the set of anchor booklets to be scored by bilingual Scorers 1 and 2.

In an ideal world, the results of these double-scoring procedures would confirm that scoring accuracy was 100% and that scorers were perfectly consistent with each



	Bundle																		
Scorer	1	2	3	4	5	6	0	7	8	9	10	11	12	13	14	15	16	17	18
1	A	B		B	A		A	A	B		B	A		A	B		B	A	
2	B		A	A		B	B	B		A	A		B	B		A	A		B
3		A	B		B	A			A	B		B	A		A	B		B	A

Fig. 9.4 Double-scoring design for PIAAC. Cells marked with "A" represent the first scorer for each bundle

other. Although this level of consistency is never obtained due to random deviations, scoring accuracy in the adult literacy surveys tends to be around 96%.

When scoring discrepancies occur, experience has shown that they fall into two distinct classes. The first type of discrepancy reveals a consistent bias on the part of one scorer, for example when one scorer is consistently more lenient than others. Because countries are required to send rescoring data for analysis at set points during the scoring process, when this situation is found, problematic scorers must be retrained or, in some cases, dismissed.

The second type of discrepancy that can be revealed through analysis of the rescoring data is more challenging to address. This occurs when the scoring results reveal general inconsistencies between the scorers, with no pattern that can be attributed to one scorer or the other. This issue has been relatively rare in the adult literacy assessments. When it has occurred, it is generally the result of a problem with an item or an error in the scoring guides. One procedure for addressing this situation includes conducting a review of all inconsistently scored responses to determine if there is a systematic pattern and, if one is found, having those items rescored. Additionally, the scoring guides for such items can be revised to clarify any issue identified as causing inconsistent scoring. When a specific problem cannot be identified and resolved, model based adjustments such as assigning unique item parameters to account for this type of country-by-item deviation may be required for one or more countries to reflect this ambiguity in scoring.

9.3.4 Statistical Procedures for Handling Missing Data

A second key methodological issue developed through experience with the large-scale literacy assessments involves the treatment of missing data due to nonresponse. Missing responses reduce the amount of information available in the cognitive assessment and thus can limit the kinds of inferences that can be made about the distribution of skills in the population based on a given set of respondents. More specifically, the relationship between skills and key background characteristics is not measured well for respondents with a high proportion of item nonresponse. This issue has been addressed in the large-scale literacy assessments by estimating conditioning coefficients based on the performance of respondents with sufficient cognitive information and applying the parameters to those respondents for whom there is insufficient performance data. This solution allows stable estimation of the model and ensures that regression of performance data on background variables is based on cases that provide sufficiently accurate information.

The two most common but least desirable ways to treat missing cases are a) to ignore them and b) to assume all missing responses can be equated to incorrect responses. Ignoring missing responses is acceptable if one can assume that missing cases occur at random and that the remaining observed cases are representative of the target population. In this case, the result would be slightly larger standard errors due to reduced sample size, and the other estimates would remain unbiased. Randomly missing data rarely occur in real data collections, however, especially in surveys of performance. If the incidence of nonresponse varies for major subgroups of interest, or if the missing responses are related to the measurement objective— in this case, the measurement of literacy skills—then inferring the missing data from observed patterns results in biased estimates. If one can be sure that all missingness is due to a lack of skill, the treatment as incorrect is justified. This treatment may be appropriate in high-stakes assessments that are consequential for respondents. In surveys, however, the respondent will not be subjected to any consequences, so other reasons for missingness, such as a lack of motivation, may be present.

To address these issues, different approaches have been developed. In order to infer reasons for nonresponse, participants are classified into two groups based on standardized coding schemes used by interviewers to record reasons for nonparticipation: those who stop the assessment for literacy-related issues (e.g., reading difficulty, native language other than language of the assessment, learning disability) and those who stop for reasons unrelated to literacy (e.g., physical disability, refusal for unspecified reason). Special procedures are used to impute the proficiencies of individuals who complete fewer than the minimum number of tasks needed to estimate their proficiencies directly.

When individuals cite a literacy-related reason for not completing the cognitive items, this implies that they were unable to respond to the items. On the other hand, citing a reason unrelated to literacy implies nothing about a person's literacy proficiency. When an individual responds to fewer than five items per scale— the minimum number needed to directly estimate proficiencies—cases are treated as follows:

- If the individual cited a literacy-related reason for not completing the assessment, then all consecutively missing responses at the end of a block of items are scored as wrong.
- If the individual cited a reason unrelated to literacy, then all consecutively missing responses at the end of block are treated as not reached.

A respondent's proficiency is calculated from a posterior distribution that is the product of two functions: a conditional distribution of proficiency, given responses to the background questionnaire; and a likelihood function of proficiency, given responses to the cognitive items (see Murray et al. 1997, for more detail). By scoring missing responses as incorrect for individuals citing literacy-related reasons for stopping the assessment, the likelihood function is very peaked at the lower end of the scale—a result that is believed to accurately represent their proficiency.

Because PIAAC was a computer-based assessment, information was available to further refine the scoring rules for non-response. The treatment of item level missing data in paper-and-pencil assessments largely has to rely on the position of items. In order to define the reason for not responding as either volitional or being based on having never been exposed to (not reached) the items, the location of the 'last' item for which a response was observed is crucial. In computer-based assessments, non-response can be treated in a more sophisticated way by taking timing data and process information into account. While the problem of rapid guessing has been described in high-stakes assessment (Wise and DeMars 2005), the nature of literacy surveys does not compel respondents to guess, but rather to skip an item rapidly for some reasons that may be unrelated to skills, for example perceived time pressure or a lack of engagement. If an item was skipped in this way – a rapid move to the next item characterized by a very short overall time spent on the item (e.g., less than 5 s) and the minimal number of actions sufficient to 'skip' the item, PIAAC applied a coding of 'not reached/not administered' (OECD 2013; Weeks et al. 2014). If, however a respondent spent time on an item, or showed more than the minimum number of actions, a missing response would be assumed to be a volitional choice and counted as not correct.

9.3.5 *Forward-Looking Design for Using Context Information in Computer-Based Assessments*

The methodologies used in large-scale assessments are well developed, and variants of essentially these same methodologies are used in all major large-scale literacy assessments. While this repeated use implies that the current methodology is well suited for the analyses of assessments at hand, new challenges have arisen with the advent of PIAAC.

As a computer-based assessment, PIAAC presents two important advantages—and challenges—when compared to earlier paper-and-pencil assessments. First is the wealth of data that a computer can provide in terms of process information. Even seemingly simple information such as knowing precisely how much time a respondent spent on a particular item can reveal important data that were never available in the paper-and-pencil assessments. The use of such data to refine the treatment of non-response data, as described above, is one example of how this information can improve measurement. Second is the opportunity to design adaptive assessments that change the selection of items depending on a respondent's performance on previous sets of items. These differences result in both new sources of information about the performance of respondents and a change in the structure of the cognitive response data given that not all test takers respond to the same set of items.

Modern psychometric methodologies are available that can improve estimation in the face of such challenges. Such methods can draw upon process and navigation data to classify respondents (Lazarsfeld and Henry 1968) with respect to the typical paths they take through scenario-based tasks, such as the ones in PIAAC's problem-solving domain. Extensions of IRT models can reveal whether this or other types of classifications exist besides the skills that respondents apply (Mislevy and Verhelst 1990; Rost 1990; von Davier and Carstensen 2007; von Davier and Rost 1995; von Davier and Yamamoto 2004; Yamamoto 1989). Additional information such as response latency can be used to generate support variables that can be used for an in-depth analysis of the validity of responses. Rapid responders (DeMars and Wise 2010) who may not provide reliable response data can potentially be identified using this data. Nonresponse models (Glas and Pimentel 2008; Moustaki and Knott 2000; Rose et al. 2010) can be used to gain a deeper understanding of situations in which certain types of respondents tend not to provide any data on at least some of the items. Elaborate response-time models that integrate latency and accuracy (Klein Entink et al. 2009; Lee 2008) can be integrated with current large-scale assessment methodologies.

9.4 Linking Real-Life Stimulus Materials and Innovative Item Types

From the first adult literacy assessment onward, items have been based on everyday materials taken from various adult situations and contexts including the workplace, community, and home. In the 1993 National Adult Literacy Survey (NALS), for example, sets of open-ended questions required respondents to use a six-page newspaper that had been created from articles, editorials, and advertisements taken from real newspapers. In PIAAC, simulation tasks were based on content from real websites, advertisements, and e-mails. For each of the large-scale literacy assessments, original materials were used in their entirety, maintaining the range of wording, formatting, and presentation found in the source. The inclusion of real-life materials both increased the content validity of the assessments and improved respondent motivation, with participants commenting that the materials were both interesting and appropriate for adults.

Each of the large-scale literacy assessments also used open-ended items. Because they are not constrained by an artificial set of response options, these open-ended tasks allowed respondents to engage in activities that are similar to those they might perform if they encountered the materials in real life. In the paper-and-pencil literacy assessments, a number of different open-ended response types were employed. These included asking respondents to underline or circle information in the stimulus, copy or paraphrase information in the stimulus, generate a response, and complete a form.

With the move to computer-based tasks in PIAAC, new ways to collect responses were required. The design for PIAAC called for the continued use of open-ended response items, both to maintain the real-life simulation focus of the assessment and to maintain the psychometric link between PIAAC and prior surveys. While the paper-and-pencil surveys allowed respondents to compose answers ranging from a word or two to several sentences, the use of automated scoring for such responses was not possible, given that PIAAC was delivered in 33 languages. Instead, the response modes used for computer-based items in this assessment included highlighting, clicking, and typing numeric responses—all of which could be scored automatically. Throughout previous paper-and-pencil assessments, there had always been some subset of respondents who marked their responses on the stimulus rather than writing answers on the provided response lines. These had been considered valid answers, and scoring rubrics had been developed to train scorers on how such responses should be scored. Thus electronic marking of text by highlighting a phrase or sentence or clicking on a cell in a table fit within existing scoring schemes. Additionally, previous work on a derivative computer-based test for individuals, the PDQ Profile Series, had shown that item parameters for paper-and-pencil items adapted from IALS and ALL were not impacted when those items were presented on the computer and respondents were asked to highlight, click, or type a numeric response. PIAAC thus became the first test to employ these response modes on a large scale and in an international context.

Taking advantage of the computer-based context, PIAAC also introduced new types of simulation items. In reading literacy, items were included that required respondents to use scrolling and hyperlinks to locate text on a website or provide responses to an Internet poll. In the new problem-solving domain, tasks were situated in simulated web, e-mail, and spreadsheet environments that contained common functionality for these environments. Examples of these new simulation tasks included items that required respondents to access information in a series of e-mails and use that information to schedule meeting rooms via an online reservation system or to locate requested information in a complex spreadsheet where the spreadsheet environment included "find" and "sort" options that would facilitate the task.

In sum, by using real-life materials and open-ended simulation tasks, ETS's large-scale literacy assessments have sought to reflect and measure the range of literacy demands faced by adults in order to provide the most useful information to policy makers, researchers, and the public. Over time, the nature of the assessment materials and tasks has been expanded to reflect the changing nature of literacy as the role of technology has become increasingly prevalent and important in everyday life.

9.5 Developing Extensive Background Questionnaires to Link Performance With Experience and Outcome Variables

One important goal of the large-scale literacy assessments has been to relate skills to a variety of demographic characteristics and explanatory variables. Doing so has allowed ETS to investigate how performance is related to social and educational outcomes and thereby interpret the importance of skills in today's society. It has also enhanced our understanding of factors related to the observed distribution of literacy skills across populations and enabled comparisons with previous surveys.

For each of the literacy assessments, respondents completed a background questionnaire in addition to the survey's cognitive measures. The background questions were a significant component of each survey, taking up to one-third of the total survey time. In each survey, the questionnaire addressed the following broad issues:

- General language background
- Educational background and experience
- Labor force participation
- Literacy activities (types of materials read and frequency of use for various purposes)
- Political and social participation
- Demographic information

As explained earlier, information collected in the background questionnaires is used in the psychometric modeling to improve the precision of the skills measurement. Equally importantly, the background questionnaires provide an extensive database that has allowed ETS to explore questions such as the following: What is the relationship between literacy skills and the ability to benefit from employer-supported training and lifelong learning? How are educational attainment and literacy skills related? How do literacy skills contribute to health and well being? What factors may contribute to the acquisition and decline of skills across age cohorts? How are literacy skills related to voting and other indices of social participation? How do reading practices affect literacy skills?

The information collected via the background questionnaires has allowed researchers and other stakeholders to look beyond simple demographic information and examine connections between the skills being measured in the assessments and important personal and social outcomes. It has also led to a better understanding of factors that mediate the acquisition or decline of skills. At ETS, this work has provided the foundation for reports that foster policy debate on critical literacy issues. Relevant reports include Kirsch et al. (2007), Rudd et al. (2004) and Sum et al. (2002, 2004).

9.6 Establishing Innovative Reporting Procedures to Better Integrate Research and Survey Data

Reports for each of the large-scale surveys have gone beyond simply reporting distributions of scores on the assessment for each participating country. As noted above, using information from the background questionnaire has made it possible to link performance to a wide range of demographic variables. Other reporting innovations have been implemented to make the survey data more useful and understandable for policy makers, researchers, practitioners, and other stakeholders.

The PIAAC data, conjointly with IALS and ALL trend data, are available in the Data Explorer (http://piaacdataexplorer.oecd.org/ide/idepiaac/), an ETS-developed web-based analysis and reporting tool that allows users to query the PIAAC database and produce tabular and graphical summaries of the data. This tool has been designed for a wide range of potential users, including those with little or no statistical background. By selecting and organizing relevant information, stakeholders can use the large-scale data to address questions of importance to them.

In addition to linking performance and background variables, survey reports have also looked at the distribution of literacy skills and how performance is related to underlying information-processing skills. Reports have included item maps that present sample items in each domain, showing where these items performed on the literacy scale and discussing features of the stimuli and questions that impact difficulty. Such analyses have allowed stakeholders to understand how items represent the construct and thereby allow them to generalize beyond the pool of items in any one assessment. These reports were also designed to provide readers with a better understanding of the information-processing skills underlying performance. Such an understanding has important implications for intervention efforts.

9.7 Conclusion

During the 30 years over which the six large-scale adult literacy assessments have been conducted, literacy demands have increased in terms of the types and amounts of information adults need to manage their daily lives. The goal of the assessments has been to provide relevant information to the variety of stakeholders interested in the skills and knowledge adults have and the impact of those skills on both individuals and society in general. Meeting such goals in this ever-changing environment has required that ETS take a leading role in the following:

- Expanding the construct of literacy
- Developing a model for building construct-based assessments
- Expanding and implementing large-scale assessment methodology to ensure reliable, valid, and comparable measurement across countries and over time
- Taking an approach to test development that focuses on the use of real-life materials and response modes that better measure the kinds of tasks adults encounter in everyday life[6]
- Developing extensive background questionnaires that make it possible to link performance with experience and outcome variables, thereby allowing the survey data to address important policy questions
- Developing reporting procedures that better integrate survey data with research

These efforts have not just expanded knowledge of what adults know and can do; they have also made important contributions to understanding how to design, conduct, and report the results of large-scale international assessments.

[6] Sample PIAAC items are available at http://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm.

Appendix: Description of the Five Levels for Prose, Document, and Numeracy Domains

	Prose	Document	Numeracy
Level 1 (0–225)	Most of the tasks in this level require the respondent to read a relatively short text to locate a single piece of information that is identical to or synonymous with the information given in the question or directive. If plausible but incorrect information is present in the text, it tends not to be located near the correct information.	Tasks in this level tend to require the respondent either to locate a piece of information based on a literal match or to enter information from personal knowledge onto a document. Little, if any, distracting information is present.	Tasks in this level require the respondent to show an understanding of basic numerical ideas by completing simple tasks in concrete, familiar contexts where the mathematical content is explicit with little text. Tasks consist of simple, one-step operations such as counting, sorting dates, performing simple arithmetic operations, or understanding common and simple percentages such as 50%.
Level 2 (226–275)	Some tasks in this level require respondents to locate a single piece of information in the text; however, several distractors or plausible but incorrect pieces of information may be present, or low-level inferences may be required. Other tasks require the respondent to integrate two or more pieces of information or to compare and contrast easily identifiable information based on a criterion provided in the question or directive.	Tasks in this level are more varied than those in level 1. Some require the respondents to match a single piece of information; however, several distractors may be present, or the match may require low-level inferences. Tasks in this level may also ask the respondent to cycle through information in a document or to integrate information from various parts of a document.	Tasks in this level are fairly simple and relate to identifying and understanding basic mathematical concepts embedded in a range of familiar contexts where the mathematical content is quite explicit and visual with few distractors. Tasks tend to include one-step or two-step processes and estimations involving whole numbers, interpreting benchmark percentages and fractions, interpreting simple graphical or spatial representations, and performing simple measurements.

(continued)

	Prose	Document	Numeracy
Level 3 (276–325)	Tasks in this level tend to require respondents to make literal or synonymous matches between the text and information given in the task, or to make matches that require low-level inferences. Other tasks ask respondents to integrate information from dense or lengthy text that contains no organizational aids such as headings. Respondents may also be asked to generate a response based on information that can be easily identified in the text. Distracting information is present but is not located near the correct information.	Some tasks in this level require the respondent to integrate multiple pieces of information from one or more documents. Others ask respondents to cycle through rather complex tables or graphs that contain information that is irrelevant or inappropriate to the task.	Tasks in this level require the respondent to demonstrate understanding of mathematical information represented in a range of different forms, such as in numbers, symbols, maps, graphs, texts, and drawings. Skills required involve number and spatial sense; knowledge of mathematical patterns and relationships; and the ability to interpret proportions, data, and statistics embedded in relatively simple texts where there may be distractors. Tasks commonly involve undertaking a number of processes to solve problems.
Level 4 (326–375)	These tasks require respondents to perform multiple-feature matches and to integrate or synthesize information from complex or lengthy passages. More complex inferences are needed to perform successfully. Conditional information is frequently present in tasks at this level and must be taken into consideration by the respondent.	Tasks in this level, like those at the previous levels, ask respondents to perform multiple-feature matches, cycle through documents, and integrate information; however, they require a greater degree of inference. Many of these tasks require respondents to provide numerous responses but do not designate how many responses are needed. Conditional information is also present in the document tasks at this level and must be taken into account by the respondent.	Tasks at this level require respondents to understand a broad range of mathematical information of a more abstract nature represented in diverse ways, including in texts of increasing complexity or in unfamiliar contexts. These tasks involve undertaking multiple steps to find solutions to problems and require more complex reasoning and interpretation skills, including comprehending and working with proportions and formulas or offering explanations for answers.

(continued)

	Prose	Document	Numeracy
Level 5 (376–500)	Some tasks in this level require the respondent to search for information in dense text that contains a number of plausible distractors. Others ask respondents to make high-level inferences or use specialized background knowledge. Some tasks ask respondents to contrast complex information.	Tasks in this level require the respondent to search through complex displays that contain multiple distractors, to make high-level text-based inferences, and to use specialized knowledge.	Tasks in this level require respondents to understand complex representations and abstract and formal mathematical and statistical ideas, possibly embedded in complex texts. Respondents may have to integrate multiple types of mathematical information, draw inferences, or generate mathematical justification for answers.

References

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*(3), 207–229. https://doi.org/10.1080/15305058.2010.496347

Glas, C. A. W., & Jehangir, K. (2014). Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 97–116). New York: Chapman & Hall

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907–922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–96). New York: Springer.

Kirsch, I. S. (2001). *The international adult literacy survey (IALS): Understanding what was measured* (Research Report No. RR-01-25). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2001.tb01867.x

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

Kirsch, I. S., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's perfect storm: Three forces changing our nation's future* (Policy Information Report). Princeton: Educational Testing Service.

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621–640. https://doi.org/10.1348/000711008X374126

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1–15. https://doi.org/10.3758/PBR.15.1.1

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 229–258). New York: Chapman & Hall.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillian.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(1), 13–23. https://doi.org/10.3102/0013189X023002013

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215. https://doi.org/10.1007/BF02295283

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14*, 147–180. https://doi.org/10.1080/01638539109544780

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163*, 445–459. https://doi.org/10.1111/1467-985X.00177

Murray, T. S., Kirsch, I. S., & Jenkins, L. B. (Eds.). (1997). *Adult literacy in OECD countries: Technical report on the first international adult literacy survey*. Washington, DC: National Center for Education Statistics.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en

OECD. (2013). Technical report of the Survey of Adult Skills (PIAAC). Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling non-ignorable missing data with IRT* (Research Report No. RR-10-11). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *AppliedPsychologicalMeasurement,3*,271–282.https://doi.org/10.1177/014662169001400305

Rudd, R., Kirsch, I., & Yamamoto, K. (2004). *Literacy and health in America* (Policy Information Report). Princeton: Educational Testing Service.

Rust, K. (2014). Sampling, weighting, and variance estimation in international large scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 117–154). New York: Chapman & Hall.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*, 142–151. https://doi.org/10.3102/0013189X10363170

Sum, A., Kirsch, I. S., & Taggart, R. (2002). *The twin challenges of mediocrity and inequality: Literacy in the U.S. from an international perspective*. Princeton: Educational Testing Service.

Sum, A., Kirsch, I. S., & Yamamoto, K. (2004). *A human capital concern: The literacy proficiency of U.S. immigrants*. Princeton: Educational Testing Service.

von Davier, M., & Carstensen, C. (Eds.). (2007). *Multivariate and mixture distribution Rasch models*. New York: Springer. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–382). New York: Springer. https://doi.org/10.1007/978-1-4612-4230-7_20

von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformation. *Methodology, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis*. Invited lecture at the ETS Spearman invitational conference, Philadelphia, PA.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (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.

Weeks, J., von Davier, M., & Yamamoto, K. (2014). Design considerations for the Programme for International Student Assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment* (pp. 259–276). New York: Chapman & Hall.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class model* (Research Report No. RR-89-41). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K. (1998). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the First International Adult Literacy Survey* (pp. 161–178). Washington, DC: National Center for Education Statistics.

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

Chapter 10
Modeling Change in Large-Scale Longitudinal Studies of Educational Growth: Four Decades of Contributions to the Assessment of Educational Growth

Donald A. Rock

ETS has had a long history of attempting to at least minimize, if not solve, many of the longstanding problems in measuring change (cf. Braun and Bridgeman 2005; Cronbach and Furby 1970; Rogosa 1995) in large-scale panel studies. Many of these contributions were made possible through the financial support of the Longitudinal Studies Branch of the U.S. Department of Education's National Center for Education Statistics (NCES). The combination of financial support from the Department of Education along with the content knowledge and quantitative skills of ETS staff over the years has led to a relatively comprehensive approach to measuring student growth. The present ETS model for measuring change argues for (a) the use of adaptive tests to minimize floor and ceiling effects, (b) a multiple-group Bayesian item response theory (IRT) approach to vertical scaling, which takes advantage of the adaptive test's potential to allow for differing ability priors both within and between longitudinal data waves, and (c) procedures for not only estimating how much an individual gains but also identifying where on the vertical scale the gain takes place. The latter concept argues that gains of equivalent size may well have quite different interpretations. The present model for change measurement was developed over a number of years as ETS's experience grew along with its involvement in the psychometric analyses of each succeeding NCES-sponsored national longitudinal study. These innovations in the measurement of change were not due solely to a small group of ETS staff members focusing on longitudinal studies, but also profited considerably from discussions and research solutions developed by the ETS NAEP group. The following historical summary recounts ETS's role in NCES's sequence of longitudinal studies and how each study contributed to the final model for measuring change.

D.A. Rock (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: donaldR706@aol.com

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_10

For the purposes of this discussion, we will define large-scale longitudinal assessment of educational growth as data collections from national probability samples with repeated and direct measurements of cognitive skills. NCES funded these growth studies in order to develop longitudinal databases, which would have the potential to inform educational policy at the national level. In order to inform educational policy, the repeated waves of testing were supplemented with the collection of parent, teacher, and school process information. ETS has been or is currently involved in the following large-scale longitudinal assessments, ordered from the earliest to the most recent:

- The National Longitudinal Study of the High School Class of 1972 (NLS-72)
- High School and Beyond (HS&B 1980–1982), sophomore and senior cohorts
- The National Education Longitudinal Study of 1988 (NELS:88)
- The Early Childhood Longitudinal Studies (ECLS):
  - Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K)
  - Early Childhood Longitudinal Study, Birth Cohort of 2001 (ECLS-B)
  - Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ECLS-K:2011)

We discuss the NLS-72 study briefly here, even though it is the only study in the list above that that does not meet one of the criteria we stated as part of our definition of large-scale, longitudinal assessment: Specifically, it does not include direct repeated cognitive measures across succeeding waves of data collection. While it was longitudinal with respect to educational attainment among post-high school participants, its shortcomings with respect to measuring change in developing cognitive skills led NCES to require the succeeding large-scale longitudinal studies to have direct repeated measures of cognitive skills. NCES and educational policy experts felt that the inclusion of repeated direct measures of cognitive skills would greatly strengthen the connection between school processes and cognitive growth. The reader should keep in mind that, while the notion of *value added* (Braun 2006) had not yet achieved its present currency, there was considerable concern about assessing the impact of selection effects on student outcomes independent of school and teaching practices. One way, or at least the first step in addressing this concern, was to measure *change* in cognitive skills during the school years. More specifically, it was hoped that measuring cognitive achievement at a relevant point in a student's development and again at a later date would help assess the impact on student growth of educational inputs and policies such as public versus private education, curriculum paths, tracking systems, busing of students across neighborhoods, and dropout rates.

As one progresses from the first to last of the above studies there was an evolutionary change in: (a) *what should be measured,* (b) *how it was measured,* and (c) *when it was measured***.** The following summary of each of the studies will detail the evolution in both ETS's and NCES's thinking in each of these three dimensions, which in the end led to ETS's most recent thinking on measuring change in cognitive skills. Obviously, as the contracting agency, NCES and its policy advisors had the final say on what was measured and when it was measured. Although ETS's main role was to provide input on development, administration, and scoring of specific cognitive measures, psychometric findings from each succeeding large-scale longitudinal assessment informed decisions with respect to all three areas. While this paper records ETS's involvement in NCES longitudinal studies, we would be remiss not to mention our partners' roles in these studies. Typically, ETS had responsibility for the development of cognitive measures and psychometric and scaling analysis as a subcontractor to another organization that carried out the other survey activities. Specifically, ETS partnered with the Research Triangle Institute (RTI) on NLS-72 and ECLS-B, the National Opinion Research Center (NORC) on HS&B, NELS-88, and Phase I of ECLS-K, and Westat on ECLS-K Phases II-IV and ECLS-K:2011.

10.1 National Longitudinal Study of 1972 (NLS-72)

NCES has referred to NLS-72 as the "grandmother of the longitudinal studies" (National Center for Education Statistics [NCES] 2011, para. 1). When the NLS-72 request for proposals was initiated, principals at NCES were Dennis Carroll, who later became the director of longitudinal studies at NCES; and William Fetters and Kenneth Stabler, NCES project managers. NCES asked bidders responding to its NLS-72 request for proposals to submit plans and budgets for sample design, data collection, and the development and scoring of the instruments. Unlike succeeding longitudinal studies, NCES awarded a single organization (ETS) the contract including base-year sample design, data collection, and instrument design and scoring; NCES did not repeat this practice on succeeding longitudinal studies. In all future bidding on longitudinal study contracts, NCES welcomed, and in fact strongly preferred, that the prime contractor not undertake all the various components alone but instead assemble consortia of organizations with specific expertise in the various survey components. We would like to think that ETS's performance on this contract had little or no bearing on the change in contracting philosophy at NCES. It was, however, true that we did not have, at the time, in-house expertise in sampling design and operational experience in collecting data on a national probability sample.

At any rate, ETS had the winning bid under the direction of Tom Hilton of the Developmental Research division and Hack Rhett from the Program Direction area. Hilton's responsibilities included insuring the alignment of the cognitive measures, and to a lesser extent the other performance measures, with the long term goals of the study. Rhett's responsibilities were primarily in the operational areas and included overseeing the data collection, data quality, and scoring of the instruments.

The primary purpose of NLS-72 was to create a source of data that researchers could use to relate student achievement and educational experiences to postsecondary educational and occupational experiences. An earlier survey of educational policymakers and researchers suggested a need for student data on educational experiences that could be related to their post-secondary occupational/educational decisions and performance. Given time and budget constraints, it was decided that a battery of cognitive measures given in the spring of the senior year could provide a reasonable summary of a student's knowledge just prior to leaving high school. Limited information about school policies and processes were gathered from a school questionnaire, a student record document, a student questionnaire, and a counselor questionnaire. Unlike succeeding NCES longitudinal studies, NLS-72 provided only indirect measures of classroom practices and teacher qualifications since there was no teacher questionnaire. Indirect measures of teacher behavior were gathered from parts of the school and student questionnaire. The base-year student questionnaire included nine sections devoted to details about the student's plans and aspirations with respect to occupational/educational decisions, vocational training, financial resources, and plans for military service. This emphasis on post-secondary planning reflected the combined interest of the stakeholders and Dennis Carroll of NCES.

Five follow-ups were eventually carried out, documenting the educational attainment and occupational status (and, in some cases, performance) of individuals sampled from the high school class of 1972. In a publication released by NCES, NLS-72 is described as "probably the richest archive ever assembled on a single generation of Americans" (NCES 2011, para. 1). The publication goes on to say, "The history of the Class of 72 from its high school years through its early 30s is widely considered as the baseline against which the progress and achievements of subsequent cohorts will be measured" (NCES 2011, para 3). ETS was not directly involved in the five follow-up data collections. The primary contractor on the five follow-ups that tracked the post-graduate activities was the Research Triangle Institute (RTI); see, for example, Riccobono et al. (1981).

The NLS-1972 base-year national probability sample included 18,000 seniors in more than 1,000 public and nonpublic schools. In the larger schools, 18 students were randomly selected while in some smaller schools all students were assessed. Schools were selected from strata in such a way that there was an over-representation of minorities and disadvantaged students. The cognitive test battery included six measures: vocabulary, mathematics, reading, picture-number associations, letter groups, and mosaic comparisons. The battery was administered in a 69-min time period. Approximately 15,800 students completed the test battery. The reader should note that the battery included three nonverbal measures: picture-number associations (rote memory), letter groups (ability to apply general concepts), and mosaic comparisons (perceptual speed and accuracy). The inclusion of nonverbal measures seemed reasonable at the time since it was believed that: (a) the oversampled disadvantaged subpopulations could be hindered on the other language-loaded measures, and (b) a mixture of aptitude and achievement measures would give a more complete picture of the skills of students entering the workforce or post-high school education. It should be kept in mind that originally the primary goal of the NLS-72 battery was to enhance the prediction of career development choices and outcomes. The three aptitude measures were from the *Kit of Factor-Referenced Tests* developed by John French while at ETS (French 1964; Ekstrom et al. 1976). Subsequent NCES longitudinal studies dropped the more aptitude-based measures and focused more on repeated achievement measures. This latter approach was more appropriate for targeting school-related gains.

Part of ETS's contractual duties included scoring the base-year test battery. No new psychometric developments (e.g., item response theory) were used in the scoring; the reported scores on the achievement tests were simply number correct. Neither NCES nor the researchers who would use the public files could be expected to be familiar with IRT procedures under development at that time. Fred Lord's seminal book on applications of item response theory (Lord 1980) was yet to appear. As we will see later, the NLS-72 achievement tests were rescored using IRT procedures in order to put them on the same scale as comparable measures in the next NCES longitudinal study: High School and Beyond (Rock et al. 1985).

NLS-72 had lofty goals:

1. Provide a national picture of post-secondary career and educational decision making.
2. Show how these decisions related to student achievement and aptitude.
3. Contrast career decisions of subpopulations of interest.

However, as in the case of all comprehensive databases, it also raised many questions. It continued to fuel the public-versus-private-school debate that Coleman (1969), Coleman and Hoffer (1987), and subsequent school effects studies initiated. Once the comparable cognitive measures for high school seniors from three cohorts, NLS-72, HS&B first follow-up (1982), and NELS:88 second follow-up (1992), were placed on the same scale, generational trends in cognitive skills could be described and analyzed. Similarly, intergenerational gap studies typically began with NLS-72 and looked at trends in the gaps between groups defined by socioeconomic status, racial or ethnic identity, and gender groups and examined how they changed from 1972 to 1992 (Konstantopoulos 2006). Researchers analyzing NLS-72 data identified additional student and teacher information that would have been helpful in describing in-school and out-of-school processes that could be related to student outcomes. Based on the experience of having discovered these informational gaps in NLS-72, NCES called for an expanded student questionnaire and the addition of a parent questionnaire in the next NCES longitudinal study, High School and Beyond, in 1980–1982.

10.2 High School and Beyond (HS&B 1980–1982)

The NCES national education longitudinal survey called High School and Beyond (HS&B) was based on a national probability sample of 10th and 12th graders (often referred to in the literature as sophomores and seniors, respectively) in the same high schools during the spring of 1980. Two years later, in 1982, the students who were 10th graders in the initial survey were re-assessed as seniors. As in the NLS-72 survey, members of the 10th grade cohort (12th graders in 1982) were followed up in order to collect data on their post-secondary activities. The HS&B sample design was a two-stage stratified cluster design with oversampling of private and Catholic schools (Frankel et al. 1981). Thirty-six students were randomly selected from the 10th and 12th grade classes in each sampled school in 1980. HS&B was designed to serve diverse users and needs while attempting to collect data reasonably comparable to NLS-72. The oversampling of private and Catholic schools allowed for specific analysis by type of school. Although multi-level analysis (Raudenbush and Bryk 2002) had not yet been formally developed, the sample of 36 students in each class made this database particularly suitable for future multi-level school effectiveness studies. That is, having 36 students in each grade significantly enhanced the reliability of the within-school regressions used in multi-level analyses later on. The significant new contributions of HS&B as contrasted to NLS-72 were:

1. The repeated testing of cognitive skills for students in their 10th grade year and then again in their 12th grade year, allowing for the measurement of cognitive development. This emphasis on the measurement of change led to a move away from a more aptitude-related test battery to a more achievement-oriented battery in subsequent surveys.
2. The use of common items shared between NLS-72 and HS&B, making possible the introduction of IRT-based common item linking (Lord 1980) that allowed intergenerational contrasts between 12th graders in NLS-72 and 12th graders in HS&B-80 in mathematics and reading.
3. The expansion of the student questionnaire to cover many psychological and sociological concepts. In the past, NCES had considered such areas too risky and not sufficiently factual and/or sufficiently researched. This new material reflected the interests of the new outside advisory board consisting of many academicians along with support from Bill Fetters from NCES. It was also consistent with awarding the HS&B base-year contract to the National Opinion Research Center (NORC), which had extensive experience in measuring these areas.
4. The introduction of a parent questionnaire administered to a subsample of the HS&B sample. The inclusion of the parent questionnaire served as both a source of additional process variables as well as a check on the reliability of student self-reports.

The primary NCES players in HS&B were Dennis Carroll, then the head of the Longitudinal Studies Branch, William Fetters, Edith Huddleston, and Jeff Owings. Fetters prepared the original survey design. The principal players among the contractors were Steve Ingels at NORC who was the prime contractor for the base year and first follow-up study. Cognitive test development and psychometrics were ETS's responsibility, led by Don Rock and Tom Hilton. Tom Donlon played a major role in the selection of the cognitive test battery, and Judy Pollack carried out psychometric analyses with the advice and assistance of Fred Lord and Marilyn Wingersky.

The final selection of the HS&B test battery did not proceed as smoothly as hoped. ETS was given the contract to revise the NLS-72 battery. The charge was to replace some of the NLS-72 tests and items and add new items, yet make the HS&B scores comparable to those of the NLS-72 battery. ETS submitted a preliminary test plan that recommended that the letter groups, picture-number associations, and mosaic comparisons subtests be dropped from the battery. This decision was made because a survey of the users of the NLS-72 data tapes and the research literature suggested that these tests were little used. Donlon et al. suggested that science and a measure of career and occupational development be added to the HS&B 10th and 12th grade batteries. They also suggested adding a spatial relations measure to the 10th grade battery and abstract reasoning to the 12th grade battery. NCES accepted these recommendations; NORC field-tested these new measures. When the field test results were submitted to the National Planning Committee for HS&B, the committee challenged the design of the batteries (cf. Heyns and Hilton 1982). The committee recommended to NCES that:

> …the draft batteries be altered substantially to allow for the measurement of school effects and cognitive change in a longitudinal framework. The concerns of the committee were twofold: First, conventional measures of basic cognitive skills are not designed to assess patterns of change over time, and there was strong feeling that the preliminary batteries would not be sufficiently sensitive to cognitive growth to allow analysis to detect differential effects among students. Second, the Committee recommended including items that would be valid measures of the skills or material a student might encounter in specific high school classes. (Rock et al. 1985, p. 27)

The batteries were then revised to make the HS&B 1980 12th grade tests a vehicle for measuring cross-sectional change from NLS-72 12th graders to HS&B 1980 12th graders. The HS&B 1980 12th grade test items were almost identical to those of NLS-72. The HS&B 1980 10th grade tests, however, were designed to be a baseline for the measurement of longitudinal change from the 10th grade to the 12th grade. The final HS&B 1980 10th grade test battery included vocabulary, reading, mathematics, science, writing, and civics education. With the possible exception of vocabulary, the final battery could be said to be more achievement-oriented than either the NLS-72 battery or the preliminary HS&B battery. The HS&B 1982 12th grade battery was identical to the HS&B-1980 10th grade battery. The purposes of the HS&B-1980 10th grade and 1982 12th grade test batteries were not just to predict post-secondary outcomes as in NLS-72, but also to measure school-related gains in achievement during the last 2 years of high school.

In 1983, NCES contracted with ETS to do a psychometric analysis of the test batteries for NLS-72 and both of the HS&B cohorts (1980 12th graders and 1980 10th graders who were 12th graders in 1982) to ensure the efficacy of:

1. Cross-sectional comparisons of NLS-72 12th graders with HS&B 12th graders.
2. The measurement of longitudinal change from the 10th grade year (HS&B 1980) to the 12th grade year (HS&B 1982).

This psychometric analysis was summarized in a comprehensive report (Rock et al. 1985) documenting the psychometric characteristics of all the cognitive measures as well as the change scores from the HS&B 1980 10th graders followed up in their 12th grade year.

ETS decided to use the three-parameter IRT model (Lord 1980) and the LOGIST computer program (Wood et al. 1976) to put all three administrations on the same scale based on common items spanning the three administrations. It is true that IRT was not necessarily required for the 10th grade to 12th grade gain-score analysis since these were identical tests. However, the crosswalk from NLS-72 12th graders to HS&B 1980 10th graders and then finally to HS&B 1982 12th graders became more problematic because of the presence of unique items, especially in the latter administration. There was one other change from NLS-72 to HS&B that argued for achieving comparability through IRT scaling, and that was the fact that NLS-72 12th graders marked an answer sheet while HS&B participants marked answers in the test booklet. As a result, HS&B test-takers attempted, on average, more items. This is not a serious problem operationally for IRT, which estimates scores based on items attempted and compensates for omitted items. Comparisons across cohorts were only done in reading and mathematics, which were present for all administrations. The IRT common crosswalk scale was carried out by pooling all test responses from all three administrations, with items not present for a particular administration treated as *not administered* for students in that particular cohort. Maximum likelihood estimates of number correct true scores were then computed for each individual.

For the longitudinal IRT scaling of the HS&B sophomore cohort tests, item parameters were calibrated separately for 10th graders and 12th graders and then transformed to the 12th grade scale. The HS&B 10th grade cohort science and writing tests were treated differently because of their shorter lengths. For the other tests, samples were used in estimating the pooled IRT parameters because the tests were sufficiently long to justify saving processing time and expense by selecting samples for item calibration. For the shorter science and writing tests, the whole sample was used.

With respect to the psychometric characteristics of the tests, it was found that:

1. The "sophomore tests were slightly more difficult than would be indicated by measurement theory" (Rock et al. 1985, p. 116). This was the compromise necessary because the same test was to be administered to 10th and 12th graders, and potential ceiling effects need to be minimized. Future longitudinal studies addressed this problem in different ways.
2. Confirmatory factor analysis (Joreskog and Sorbom 1996) suggested that the tests were measuring the same things with the same precision across racial/ethnic and gender groups.
3. Traditional estimates of reliability increased from the 10th grade to the 12th grade year in HS&B. Reliability estimates for IRT scores were not estimated. Reliability of IRT scores, however, would be estimated in subsequent longitudinal studies.
4. While the psychometric report argues that mathematics, reading, and science scores were sufficiently reliable for measuring individual change, they were borderline by today's criteria. Most of the subtests, with about 20 items each, had alpha coefficients between .70 and .80. The mathematics test, with 38 items, had

> alpha coefficients close to .90 for the total group and most subgroups in both years, while the civics education subtest, with only 10 items, had reliabilities in the .50s, and was considered to be too low for estimating reliable individual change scores.

The HS&B experience taught us a number of lessons with respect to test development and methodological approaches to measuring change. These lessons led to significant changes in how students were tested in subsequent large-scale longitudinal studies. In HS&B, each student was administered six subject tests during a 69-min period, severely limiting the number of items that could be used, and thus the tests' reliabilities. Even so, there were those on the advisory committee who argued for subscores in mathematics and science. The amount of classroom time that schools would allow outside entities to use for testing purposes was shrinking while researchers and stakeholders on advisory committees increased their appetites for the number of things measured. NAEP's solution to this problem, which was just beginning to be implemented in the early 1980s, was to use sophisticated Bayesian algorithms to shrink individual scores towards their subgroup means, and then restrict reporting to summary statistics such as group means. The longitudinal studies approach has been to change the type of test administration in an attempt to provide individual scores that are sufficiently reliable that researchers can relate educational processes measured at the individual level with individual gain scores and/or gain trajectories. That is, ETS's longitudinal researchers' response to this problem was twofold: measure fewer things in a fixed amount of time, and develop procedures for measuring them more efficiently. ETS suggested that an adaptive test administration can help to increase efficiency by almost a factor of 2. That is, the IRT information function from an adaptive test can approximate that of a linear test twice as long. That is what ETS proposed for the next NCES longitudinal study.

ETS's longitudinal researchers also learned that maximum likelihood estimation (MLE) of item parameters and individual scores has certain limitations. Individuals with perfect or below-chance observed scores led to boundary condition problems, with the associated estimates of individual scores going to infinity. If we were to continue to use MLE estimation procedures, an adaptive test could help to minimize the occurrence of these problematic perfect and below-chance scores.

It is also the case that when the IRT procedures described in Lord (1980) first became popular, many applied researchers, policy stakeholders, members of advisory committees, and others got the impression that the weighted scoring in IRT would allow one to gather more reliable information in a shorter test. The fact was that solutions became very computationally unstable as the number of items became fewer in MLE estimation as used in the popular IRT program LOGIST (Wood et al. 1976). It was not until Bayesian IRT methods (Bock and Aiken 1981; Mislevy and Bock 1990) became available that stable solutions to IRT parameter estimation and scoring were possible for relatively short tests.

There is one other misconception that seems to be implicit, if not explicit, in thinking about IRT scoring—that is, the impression that IRT scores have the property of equal units along the score scale. This would be very desirable for the interpretation of gain scores. If this were the case, then a 2-point gain at the top of the test score scale would have a similar meaning with respect to progress as a 2-point gain at the bottom of the scale. This is the implicit assumption when gain scores from different parts of the test score scale are thrown in the same pool and correlated with process variables. For example, why would one expect a strong positive correlation between the number of advanced mathematics courses and this undifferentiated pool of mathematics gains? Gains at the lower end of the scale indicate progress in basic mathematics concepts while gains of an equivalent number of points at the top of the scale suggest progress in complex mathematical solutions. Pooling individual gains together and relating them to processes that only apply to gains at particular locations along the score scale is bound to fail and has little or nothing to do with the reliability of the gain scores. Policy makers who use longitudinal databases in an attempt to identify processes that lead to gains need to understand this basic measurement problem. Steps were taken in the next longitudinal study to develop measurement procedures to alleviate this concern.

10.3 The National Education Longitudinal Study of 1988 (NELS:88)

A shortcoming of the two longitudinal studies described above, NLS:72 and HS&B, is that they sampled students in their 10th or 12th-grade year of high school. As a result, at-risk students who dropped out of school before reaching their 10th or 12th-grade year were not included in the surveys. The National Education Longitudinal Study of 1988 (NELS:88) was designed to address this issue by sampling eighth graders in 1988 and then monitoring their transitions to later educational and occupational experiences. Students received a battery of tests in the eighth grade base year, and then again 2 and 4 years later when most sample members were in 10th and 12th grades. A subsample of dropouts was retained and followed up. Cognitive tests designed and scored by ETS were included in the first three rounds of data collection, in 1988, 1990, and 1992, as well as numerous questionnaires collecting data on experiences, attitudes, and goals from students, schools, teachers, and parents. Follow-ups conducted after the high school years as the students progressed to post-secondary education or entered the work force included questionnaires only, not cognitive tests. Transcripts collected from the students' high schools also became a part of this varied archive.

NELS:88 was sponsored by the Office of Educational Research and Improvement of the National Center for Education Statistics (NCES). NELS:88 was the third longitudinal study in the series of longitudinal studies supported by NCES and in which ETS longitudinal researchers participated. ETS's bidding strategy for the NELS:88 contract was to write a proposal for the test development, design of the testing procedure, and scoring and scaling of the cognitive tests. ETS's proposal was submitted as a subcontract with each of the competing prime bidders' proposals.

ETS continued to follow this bidding model for the next several longitudinal studies. Regardless of whom the prime contractor turned out to be, this strategy led to ETS furnishing considerable continuity, experience, and knowledge to the measurement of academic gain. The National Opinion Research Center (NORC) won the prime contract, and ETS was a subcontractor to NORC. Westat also was a subcontractor with responsibility for developing the teacher questionnaire. The contract monitors at NCES were Peggy Quinn and Jeff Owings, while Steven Ingels and Leslie Scott directed the NORC effort. Principals at ETS were Don Rock and Judy Pollack, aided by Trudy Conlon and Kalle Gerritz in test development. Kentaro Yamamoto at ETS also contributed very helpful advice in the psychometric area.

The primary purpose of the NELS:88 data collection was to provide policy-relevant information concerning the effectiveness of schools, curriculum paths, special programs, variations in curriculum content and exposure, and/or mode of delivery in bringing about educational growth (Rock et al. 1995; Scott et al. 1995). New policy-relevant information was available in NELS:88 with the addition of teacher questionnaires that could be directly connected with individual students. For the first time, a specific principal questionnaire was also included. Grades and course-taking history were collected in transcripts provided by the schools for a subset of students.

While the base-year (1988) sample consisted of 24,599 eighth graders, the first and second follow-up samples were smaller. As the base-year eighth graders moved on to high school, some high schools had a large number of sampled students, while others had only one or two. It would not have been cost effective to follow up on every student, which would have required going to thousands of high schools. Instead of simply setting a cutoff for retaining individual participants (e.g., only students in schools with at least ten sample members), individuals were followed up with varying probabilities depending on how they were clustered within schools. In this way, the representativeness of the sample could be maintained.

ETS test development under Trudy Conlon and Kalle Gerritz assembled an eighth-grade battery consisting of the achievement areas of reading comprehension, mathematics, science, and history/citizenship/geography. The battery was designed to measure school-related growth spanning a 4-year period during which most of the participants were in school. The construction of the NELS:88 eighth-grade battery was a delicate balancing act between several competing objectives—for example, general vs. specific knowledge and basic skills vs. higher-order thinking and problem solving. In the development of NELS:88 test items, efforts were made to take a middle road in the sense that our curriculum experts were instructed to select items that tapped the general knowledge that was found in most curricula but that typically did not require a great deal of isolated factual knowledge. The emphasis was to be on understanding concepts and measuring problem-solving skills (Rock and Pollack 1991; Ingels et al. 1993). However, it was thought necessary also to assess the basic operational skills (e.g., simple arithmetic and algebraic operations), which are the foundations for successfully carrying out the problem-solving tasks.

This concern with respect to developing tests that are sensitive to changes resulting from school related processes is particularly relevant to measuring change over relatively long periods of exposure to varied educational treatments. That is, the 2-year gaps between retesting coupled with a very heterogeneous student population were likely to coincide with considerable variability in course taking experiences. This fact, along with the constraints on testing time, made coverage of specific curriculum-related knowledge very difficult. Also, as indicated above, specificity in the knowledge being tapped by the cognitive tests could lead to distortions in the gain scores due to forgetting of specific details. The impact on gain scores due to forgetting should be minimized if the cognitive battery increasingly emphasizes general concepts and development of problem-solving abilities. This emphasis should increase as one goes to the tenth and twelfth grades. Students who take more high-level courses, regardless of the specific course content, are likely to increase their conceptual understanding as well as gain additional practice in problem-solving skills.

At best, any nationally representative longitudinal achievement testing program must attempt to balance testing-time burdens, the natural tensions between local curriculum emphasis and more general mastery objectives, and the psychometric constraints (in the case of NELS:88 in carrying out both vertical equating [year-to-year] and cross-sectional equating [form-to-form within year]). NELS:88, fortunately, did have the luxury of being able to gather cross-sectional pretest data on the item pools. Thus, we were able to take into consideration not only the general curriculum relevance but also whether or not the items demonstrated reasonable growth curves, in addition to meeting the usual item analysis requirements for item quality.

Additional test objectives included:

1. There should be little or no floor or ceiling effects. Tests should give every student the opportunity to demonstrate gain: some at the lower end of the scale and others making gains elsewhere on the scale. As part of the contract, ETS developed procedures for sorting out where the gain takes place.
2. The tests should be unspeeded.
3. Reliabilities should be high and the standard error of measurement should be invariant across ethnic and gender groups.
4. The comparable tests should have sufficient common items to provide crosswalks to HS&B tests.
5. The mathematics test should share common items with NAEP to provide a crosswalk to NAEP mathematics.
6. If psychometrically justified, the tests should provide subscale scores and/or proficiency levels, yet be sufficiently unidimensional as to be appropriate for IRT vertical scaling across grades.
7. The test battery should be administered within an hour and a half.

Obviously, certain compromises needed to be made, since some of the constraints are in conflict. In order to make the test reliable enough to support change-measurement within the time limits, adaptive testing had to be considered. It was decided that two new approaches would be introduced in the NELS:88 longitudinal study.

The first approach was the introduction of multi-stage adaptive testing (Cleary et al. 1968; Lord 1971) in Grade 10 and Grade 12. Theoretically, using adaptive tests would maximize reliability (i.e., maximize the expected IRT information function) across the ability distribution and do so with fewer items. Even more importantly, it would greatly minimize the potential for having floor and ceiling effects, the bane of all gain score estimations.

The second innovation was the identification of clusters of items identifying multiple proficiency levels marking a hierarchy of skill levels on the mathematics, reading comprehension, and science scales. These proficiency levels could be interpreted in much the same way as NAEP's proficiency levels, but they had an additional use in measuring gain: They could be used to pinpoint where on the scale the gain was taking place. Thus, one could tell not only *how much* a given student gained, but also *at what skill level* he or she was gaining. This would allow researchers and policymakers to select malleable factors that could influence gains at specific points (proficiency levels) on the scale. In short, this allowed them to match the educational process (e.g., taking a specific course), with the location on the scale where the maximum gain would be expected to be taking place.[1]

10.3.1 The Two-Stage Multilevel Testing in the NELS:88 Longitudinal Framework

The potentially large variation in student growth trajectories over a 4-year period argued for a longitudinal tailored testing approach to assessment. That is, to accurately assess a student's status both at a given point in time as well as over time, the individual tests must be capable of measuring across a broad range of ability or achievement. In the eighth-grade base year of NELS:88, all students received the same test battery, with tests designed to have broadband measurement properties. In the subsequent years, easier or more difficult reading and mathematics forms were selected according to students' performance in the previous years. A two-stage multilevel testing procedure was implemented that used the eighth-grade reading and mathematics test score results for each student to assign him or her to one of two forms in 10th-grade reading, and one of three forms in 10th grade mathematics, that varied in difficulty. If the student did very well (top 25%) on the eighth-grade

[1] The concept that score gains at different points on the scale should (a) be interpreted differently and (b) depending on that interpretation, be related to specific processes that affect that particular skill, has some intellectual forebears. For example, Cronbach and Snow (1977) described the frequent occurrence of aptitude-by-treatment interaction in educational pre-post test designs. We would argue that what they were observing was the fact that different treatments were necessary because they were looking for changes along different points on the aptitude scale. From an entirely different statistical perspective, Tukey, in a personal communication, once suggested that most if not all interactions can be reduced to nonsignificance by applying the appropriate transformations. That may be true operationally, but we might be throwing away the most important substantive findings.

mathematics test, he or she received the most difficult of the three mathematics forms in 10th grade; conversely, students scoring in the lowest 25% received the easiest form 2 years later. The remaining individuals received the middle form. With only two reading forms to choose from in the follow-up, the routing cut was made using the median of the eighth-grade scores. This branching procedure was repeated 2 years later, using 10th-grade performance to select the forms to be administered in 12th grade.

The 10th- and 12th-grade tests in reading and mathematics were designed to include sufficient linking items across grades, as well as across forms within grade, to allow for both cross-sectional and vertical scaling using IRT models. Considerable overlap between adjacent second-stage forms was desirable to minimize the loss of precision in case of any misassignment. If an individual were assigned to the most difficult second-stage form when he or she should have been assigned to the easiest form, then that student would not be well assessed, to say the least. Fortunately, we found no evidence for such two-level misclassifications. The science and history/citizenship/geography tests used the same relatively broad-ranged form for all students; linking items needed to be present only across grades.

To take advantage of this modest approach to paper-and-pencil adaptive testing, more recent developments in Bayesian IRT procedures (Mislevy and Bock 1990; Muraki and Bock 1991) were implemented in the first IRT analysis. The Bayesian procedures were able to take advantage of the fact that the adaptive procedure identified subpopulations, both within and across grades, who were characterized by different ability distributions. Both item parameters and posterior means were estimated for each individual at each point in time using a multiple-group version of PARSCALE (Muraki and Bock 1991), with updating of normal priors on ability distributions defined by grade and form within grade. PARSCALE does allow the shape of the priors to vary, but we have found that the smoothing that came from updating with normal ability priors leads to less jagged looking posterior ability distributions and does not over-fit items. It was our feeling that, often, lack of item fit was being absorbed in the shape of the ability distribution when the distribution was free to be any shape.

This procedure required the pooling of data as each wave was completed. This pooling often led to a certain amount of consternation at NCES, since item parameters and scores from the previous wave were updated as each new wave of data became available. In a sense, each wave of data remade history. However, this pooling procedure led to only very minor differences in the previous scores and tended to make the vertical scale more internally consistent. In most cases, it is best to use all available information in the estimation, and this use is particularly true in longitudinal analysis where each additional wave adds new supplementary information on item parameters and individual scores. The more typical approach fixes the linking item parameter values from the previous wave, but this procedure tends to underestimate the score variances in succeeding waves, contributing to the typical finding of a high negative correlation between initial status and gain.

It should be kept in mind that the multiple-group PARSCALE finds those item parameters that maximize the likelihood across all groups (in this case, forms):

seven in mathematics (one base-year form; three alternative forms in each follow-up), five in reading (two alternative forms per follow-up), and three each in science and history/citizenship/geography (one form per round). The version of the multiple-group PARSCALE used at that time only saved the subpopulation means and standard deviations and not the individual expected a posteriori (EAP) scores. The individual EAP scores, which are the means of their posterior distributions of the latent variable, were obtained from the NAEP B-group conditioning program, which uses the Gaussian quadrature procedure. This variation is virtually equivalent to conditioning (e.g., see Mislevy et al. 1992, as well as Barone and Beaton, Chap. 8, and Kirsch et al., Chap. 9, in this volume) on a set of dummy variables defining from which ability subpopulation an individual comes.

In summary, this procedure finds the item parameters that maximize the likelihood function across all groups (forms and grades) simultaneously. The items can be put on the same vertical scale because of the linking items that are common to different forms across years, or adjacent forms within year. Using the performance on the common items, the subgroup means can be located along the vertical scale. Individual ability scores are not estimated in the item parameter estimation step; only the subgroup means and variances are estimated. Next, NAEP's B-group program was used to estimate the individual ability scores as the mean of an individual's posterior distribution. (A detailed technical description of this procedure may be found in Rock et al. 1995). Checks on the goodness of fit of the IRT model to the observed data were then carried out.

Item traces were inspected to ensure a good fit throughout the ability range. More importantly, estimated proportions correct by item by grade were also estimated in order to ensure that the IRT model was both reproducing the item P-plus values and that there was no particular bias in favor of any particular grade. Since the item parameters were estimated using a model that maximizes the goodness-of-fit across the subpopulations, including grades, one would not expect much difference here. When the differences were summed across all items for each test, the maximum discrepancy between observed and estimated proportion correct for the whole test was .7 of a scale score point for Grade 12 mathematics, whose score scale had a range of 0 to 81. The IRT estimates tended to slightly underestimate the observed proportions. However, no systematic bias was found for any particular grade.

10.3.2 Criterion-Referenced Proficiency Levels

In addition to the normative interpretations in NELS:88 cognitive tests, the reading, mathematics, and science tests also provided criterion-referenced interpretations. The criterion-referenced interpretations were based on students demonstrating proficiencies on clusters of four items that mark ascending points on the test score scale. For example, there are three separate clusters consisting of four items each in reading comprehension that mark the low, middle, and high end of the reading scale. The items that make up these clusters exemplify the skills required to successfully answer the typical item located at these points along the scale. There were three levels in the reading comprehension test, five in the mathematics test, and three in the science test. Specific details of the skills involved in each of the levels may be found in Rock et al. (1995).

10.3.3 Criterion-Referenced Scores

There were two kinds of criterion-referenced proficiency scores reported in NELS:88 dichotomous scores and probability scores.

In the case of a dichotomous score, a 1 indicates mastery of the material in a given cluster of items marking a point on the scale, while a 0 implies nonmastery. A student was defined to be proficient at a given proficiency level if he or she got at least three out of four items correct that marked that level. Items were selected for a proficiency level if they shared similar cognitive processing demands and this cognitive demand similarity was reflected in similar item difficulties. Test developers were asked to build tests in which the more difficult items required all the skills of the easier items plus at least one additional higher level skill. Therefore, in the content-by-process test specifications, variation in item difficulty often coincided with variation in process. This logic leads to proficiency levels that are hierarchically ordered in the sense that mastery of the highest level among, for example, three levels implies that one would have also mastered the lower two levels. A student who mastered all three levels in reading had a proficiency score pattern of [1 1 1]. Similarly, a student who had only mastered the first two levels, but failed to answer at least three correct on the third level, had a proficiency score pattern of [1 1 0]. Dichotomous scores were not reported for students who omitted items that were critical to determining a proficiency level or who had reversals in their proficiency score pattern (a failed level followed by a passed level, such as 0 0 1). The vast majority of students did fit the hierarchical model; that is, they had no reversals.

Analyses using the dichotomous proficiency scores included descriptive statistics that showed the percentages of various subpopulations who demonstrated proficiencies at each of the hierarchical levels. They can also be used to examine patterns of change with respect to proficiency levels. An example of descriptive analysis using NELS:88 proficiency levels can be found in Rock et al. (1993).

The second kind of proficiency score is the probability of being proficient at each of the levels. These probabilities were computed using all of the information provided by students' responses on the whole test, not just the four-item clusters that marked the proficiency levels. After IRT calibration of item parameters and student ability estimates (thetas had been computed), additional *superitems* were defined marking each of the proficiency levels. These superitems were the dichotomous scores described above. Then, holding the thetas fixed, item parameters were calibrated for each of the superitems, just as if they were single items. Using these item parameters in conjunction with the students' thetas, probabilities of proficiency were computed for each proficiency level.

The advantages of the probability of being proficient at each of the levels over the dichotomous proficiencies are that (a) they are continuous scores and thus more powerful statistical methods may be applied, and (b) probabilities of being proficient at each of the levels can be computed for any individual who had a test score in a given grade, not only the students who answered enough items in a cluster. The latter advantage is true since the IRT model enables one to estimate how students would perform on those items that they were not given, for example, if the items were on a different form or not given in that grade.

The proficiency probabilities are particularly appropriate for relating specific processes to changes that occur at different points along the score scale. For example, one might wish to evaluate the impact of taking advanced mathematics courses on changes in mathematics achievement from Grade 10 to Grade 12. One approach to doing this evaluation would be to subtract every student's 10th-grade IRT-estimated number right from his or her 12th grade IRT-estimated number right and correlate this difference with the number of advanced mathematics courses taken between the 10th and 12th grades. The resulting correlation will be relatively low because lower achieving individuals taking no advanced mathematics courses are also gaining, *but probably at the low end of the test score scale.* Individuals who are taking advanced mathematics courses are making their greatest gains at the higher end of the test score scale. To be more concrete, let us say that the individuals who took none of the advanced math courses gained, on average, three points, all at the low end of the test score scale. Conversely, the individuals who took the advanced math courses gained three points, but virtually all of these individuals made their gains at the upper end of the test score scale. When the researcher correlates number of advanced courses with gains, the fact that, on average, the advanced math takers gained the same amount as those taking no advanced mathematics courses will lead to a very small or zero correlation between gain and specific processes (e.g., advanced math course taking). This low correlation has nothing to do with reliability of gain scores, but it has much to do with where on the test score scale the gains are taking place. Gains in the upper end of the test score distribution reflect increases in knowledge in advanced mathematical concepts and processes while gains at the lower end reflect gains in basic arithmetical concepts. In order to successfully relate specific processes to gains, one has to match the process of interest to where on the scale the gain is taking place.

The proficiency probabilities do this matching because they mark ascending places on the test score scale. If we wish to relate the number of advanced math courses taken to changes in mathematics proficiency, we should look at changes at the upper end of the test score distribution, not at the lower end, where students are making progress in more basic skills. There are five proficiency levels in mathematics, with Level 4 and Level 5 marking the two highest points along the test score scale. One would expect that taking advanced math courses would have its greatest impacts on changes in probabilities of being proficient at these highest two levels. Thus, one would simply subtract each individual's tenth grade probability of being

Table 10.1 Reliability of theta

	Baseyear	First follow-up	Second follow-up
Reading	.80	.86	.85
Math	.89	.93	.94
Science	.73	.81	.82
History/citizenship/geography	.84	.85	.85

proficient at, say, Level 4 from the corresponding probability of being proficient at Level 4 in 12th grade. Now, every individual has a continuous measure of change in mastery of advanced skills, not just a broadband change score. If we then correlate this change in Level 4 probabilities with the number of advanced mathematics courses taken, we will observe a substantial increase in the relationship between change and process (number of advanced mathematics courses taken) compared with change in the broad-band measure. We could do the same thing with the Level 5 probabilities as well. The main point here is that certain school processes, in particular course-taking patterns, target gains at different points along the test score distribution. It is necessary to match the type of school process we are evaluating with the location on the test score scale where the gains are likely to be taking place and then select the proper proficiency levels for appropriately evaluating that impact. For an example of the use of probability of proficiency scores to measure mathematics achievement gain in relation to program placement and course taking, see Chapter 4 of Scott et al. (1995).

10.3.4 Psychometric Properties of the Adaptive Tests Scores and the Proficiency Probabilities Developed in NELS:88

This section presents information on the reliability and validity of the adaptive test IRT (EAP) scores as well as empirical evidence of the usefulness of the criterion-referenced proficiency probabilities in measuring change. Table 10.1 presents the reliabilities of the thetas for the four tests. As expected, the introduction of the adaptive measures in Grades 10 and 12 lead to substantial increases in reliability. These IRT-based indices are computed as 1 minus the ratio of the average measurement error variance to the total variance.

The ETS longitudinal researchers moved from MLE estimation using LOGIST to multigroup PARSCALE and finally to NAEP's B-Group conditioning program for EAP estimates of theta and number-right true scores. The B-Group conditioning was based on ability priors associated with grade and test form. A systematic comparison was carried out among these competing scoring procedures. One of the reasons for introducing adaptive tests and Bayesian scoring procedures was to increase the accuracy of the measurement of gain by reducing floor and ceiling effects and thus enhance the relationships of test scores with relevant policy variables.

Table 10.2 Evaluation of alternative test scoring procedures for estimating gains in mathematics and their relationship with selected background/policy variables

Gains in theta metric	Any math last 2 years	Taking math now	Curriculum acad = 1; Gen/Voc = 0
Gain 8–10 LOG	0.07	0.06	0.06
Gain 8–10 STI	0.11	0.11	0.15
Gain 8–10 ST4	0.08	0.06	0.07
Gain 10–12 LOG	0.07	0.15	0.06
Gain 10–12 ST1	0.14	0.23	0.14
Gain10–12 ST4	0.10	0.18	0.06
Total gain LOG	0.12	0.18	0.11
Total gain ST1	0.19	0.26	0.22
Total gain ST4	0.14	0.18	0.10

Note. LOG = LOGIST, ST1 = NALS 1-step, ST4 = NAEP 4-step method

Table 10.3 Correlations between gains in proficiency at each mathematics level and mathematics course taking (no. of units), average grade, and precalculus course-taking

8th–12th grade gains in proficiency/ probabilities at each level in math	No. of units	Average grade	Precalculus Yes = 1; No = 0
Math level 1	−0.26	−0.28	−0.20
Math level 2	−0.01	−0.20	−0.20
Math level 3	0.22	0.05	−0.02
Math level 4	0.44	0.46	0.29
Math level 5	0.25	0.38	0.33

Table 10.2 presents a comparison of the relationships between MLE estimates and two Bayesian estimates with selected outside policy variables.

Inspection of Table 10.2 indicates that in the theta metric, the normal prior Bayesian procedure (ST1) shows stronger relationships between gains and course-taking than do the other two procedures. The differences in favor of ST1 are particularly strong where contrasts are being made between groups quite different in their mathematics preparation, for example, the relationship between being in the academic curriculum or taking math now and total gain.

When the correlations are based on the *number correct true score metric* (NCRT), the ST1 Bayesian approach still does as well or better than the other two approaches. The NCRT score metric is a nonlinear transformation of the theta scores, computed by adding the probabilities of a correct answer for all items in a selected item pool. Unlike the theta metric, the NCRT metric does not stretch out the tails of the score distribution. The stretching out at the tails has little impact on most analyses where group means are used. However, it can distort gain scores for individuals who are in or near the tails of the distribution. Gains in proficiency probabilities at each proficiency level and their respective correlations with selected process variables are shown in Table 10.3. The entries in Table 10.3 demonstrate the importance of relating specific processes with changes taking place at appropriate points along the score distribution.

Inspection of Table 10.3 indicates that gains between 8th and 12th grade in the probability of being proficient at Level 4 show a relatively high positive correlation with number of units of mathematics (.44) and with average grade in mathematics (.46). The changes in probability of mastery at each mathematics level shown in Table 10.3 are based on the ST1 scoring system.

When the dummy variable contrasting whether an individual took precalculus courses was correlated with gains in probabilities at the various proficiency levels, one observes negative correlations for demonstrated proficiencies at the two lower levels (simple operations and fractions and decimals) and higher positive correlations for Levels 4–5. That is, individuals with a score of 1 on the dummy variable, indicating they took precalculus courses, are making progressively greater gains in probabilities associated with mastery of Levels 4–5. As another example of the relation between scale region and educational process, students in the academic curriculum versus the general/vocational curriculum tend to have high positive correlations with changes in proficiency probabilities marking the high end of the scale. Conversely, students in the general/vocational curriculum tend to show positive correlations with gains in proficiency probabilities marking the low end of the scale. Other patterns of changes in lower proficiency levels and their relationship to appropriate process variables may be found in Rock et al. (1985).

10.3.5 Four New Approaches in Longitudinal Research

What did the ETS longitudinal studies group learn from NELS:88? Four new approaches were introduced in this longitudinal study. First, it was found that even a modest approach to adaptive testing improved measurement throughout the ability range and minimized floor and ceiling effects. Improved measurement led to significantly higher reliabilities as the testing moved from the 8th grade to more adaptive procedures in the 10th and 12th grades. Second, the introduction of the Bayesian IRT methodology with separate ability priors on subgroups of students taking different test forms, and/or in different grades, contributed to a more well-defined separation of subgroups both across and within grades. Third, on the advice of Kentaro Yamomoto, it became common practice in longitudinal research to pool and update item parameters and test scores as each succeeding wave of data was added. This pooling led to an internally consistent vertical scale across testing administrations. Last, we developed procedures that used criterion-referenced points to locate where on the vertical scale an individual was making his or her gains. As a result, the longitudinal researcher would have two pieces of information for each student: how much he or she gained in overall scale score points and where on the scale the gain took place. Changes in probabilities of proficiency at selected levels along the vertical scale could then be related to the appropriate policy variables that reflect learning at these levels.

While the above psychometric approaches contributed to improving longstanding problems in the measurement of change, there was still room for improvement.

For example, real-time two-stage adaptive testing would be a significant improvement over that used in the NELS:88 survey, where students' performance 2 years earlier was used to select test forms. Such an approach would promise a better fit of item difficulties to a student's ability level. This improvement would wait for the next NCES longitudinal study: The Early Childhood Longitudinal Study - Kindergarten Class of 1998–1999 (ECLS-K).

10.4 Early Childhood Longitudinal Study—Kindergarten Class of 1998–1999 (ECLS-K)

The Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) was sponsored by NCES and focused on children's school and home experiences beginning in fall kindergarten and continuing through 8th grade. Children were assessed in the fall and spring of kindergarten (1998–1999), the fall and spring of 1st grade (1999–2000), the spring of 3rd grade (2002), the spring of 5th grade (2004), and finally spring of 8th grade (2007). This was the first time that a national probability sample of kindergartners was followed up with repeated cognitive assessments throughout the critical early school years. ETS's longitudinal studies group continued the bidding strategy of writing the same psychometric proposal for inclusion in all the proposals of the prime contract bidders. NORC won the contract to develop instruments and conduct field tests prior to the kindergarten year; Westat was the winning bidder for the subsequent rounds, with ETS subcontracted to do the test development, scaling, and scoring. This study was by far the most complex as well as the largest undertaking to date with respect to the number and depth of the assessment instruments.

The spanning of so many grades with so many instruments during periods in which one would expect accelerated student growth complicated the vertical scaling. As a result, a number of subcontracts were also let reflecting the individual expertise required for the various instruments. Principals at NCES were Jeff Owings, the Longitudinal Studies Branch chief, with Jerry West, and later, Elvira Germino Hausken as project directors. The Westat effort was led by Karen Tourangeau, while NORC was represented by Tom Hoffer, who would be involved in student questionnaire construction, and Sally Atkins-Burnett and Sam Meisels from the University of Michigan led the development of indirect measures of socio-emotional and cognitive achievement. At ETS, Don Rock, Judy Pollack, and in the later rounds, Michelle Najarian, led the group responsible for developing and selecting test items and for scaling and scoring the direct measures of cognitive development. The test development endeavor benefited from the help and advice of the University of Michigan staff.

The ECLS-K base-year sample was a national probability sample of about 22,000 children who had entered kindergarten either full-day or part-day in fall 1998. About 800 public schools and 200 private schools were represented in the sample. Children in the kindergarten through fifth-grade rounds were assessed individually using computer-assisted interviewing methods, while group paper-and-pencil assessments were conducted in the eighth grade.[2] Children in the early grades (K-1) were assessed with socio-emotional and psychomotor instruments and ratings of cognitive development as well as direct cognitive assessments (Adkins-Burnett et al. 2000). The direct cognitive assessment in K-1 included a battery consisting of reading, mathematics, and general knowledge, all of which were to be completed in 75 min, on average, although the tests were not timed. In Grade 3, the general knowledge test was dropped and replaced with a science test. The original NCES plan was to assess children in fall and spring of their kindergarten year, fall and spring of their first-grade year, and in the spring only of each of their second- through fifth-grade years. Unfortunately, NCES budgetary constraints resulted in the second- and fourth-grade data collections being dropped completely; for similar reasons, data was collected from a reduced sample in fall of the first-grade year. At a later time, high school assessments were planned for 8th, 10th, and 12th grades, but again, due to budget constraints, only the 8th-grade survey was conducted.

Gaps of more than a year in a longitudinal study during a high-growth period can be problematic for vertical scaling. Dropping the second-grade data collection created a serious gap, particularly in reading. Very few children finish first-grade reading fluently; most are able to read with comprehension by the end of third grade. With no data collection bridging the gap between the early reading tasks of the first grade assessment and the much more advanced material in the third grade tests, the development of a vertical scale was at risk. As a result, a bridge study was conducted using a sample of about 1000 second graders; this study furnished the linking items to connect the first grade with the third grade and maintain the vertical scale's integrity. Subsequent gaps in data collection, from third to fifth grade and then to eighth grade were less serious because there was more overlap in the ability distributions.

While the changes referred to above did indeed complicate IRT scaling, one large difference between ECLS-K and the previous high school longitudinal studies was the relative uniformity of the curricula in the early grades. This standardization

[2] The individually administered test approach used in kindergarten through fifth grade had both supporters and critics among the experts. Most felt that individual administration would be advantageous because it would help maintain a high level of motivation in the children. In general, this was found to be true. In the kindergarten and first-grade rounds, however, some expressed a concern that the individual mode of administration may have contributed unwanted sources of variance to the children's performance in the direct cognitive measures. Unlike group administrations, which in theory are more easily standardized, variance attributable to individual administrators might affect children's scores. A multilevel analysis of fall-kindergarten and spring-first grade data found only a very small interviewer effect of about 1–3% of variance. A team leader effect could not be isolated, because it was almost completely confounded with primary sampling unit. Analysis of interviewer effect was not carried out for subsequent rounds of data for two reasons. First, the effect in kindergarten through first grade was about twice as large for the general knowledge assessment (which was not used beyond kindergarten) than for reading or mathematics. Second, the effect found was so small that it was inconsequential. Refer to Rock and Pollack (2002b) for more details on the analysis of interviewer effects.

holds reasonably well all the way through to the fifth grade. This curricular standardization facilitated consensus among clients, test developers, and outside advisors on the test specifications that would define the pools of test items that would be sensitive to changes in a child's development. However, there were some tensions with respect to item selection for measuring change across grades. While the curriculum experts emphasized the need for grade-appropriate items for children in a given grade, it is precisely the nongrade-appropriate items that also must be included in order to form links to the grade above and the grade below. Those items serve not only as linking items but also play an important role in minimizing floor and ceiling effects. Grade-appropriate items play a larger role in any cross-sectional assessment, but are not sufficient for an assessment in a particular grade as part of an ongoing longitudinal study.

Many of the psychometric approaches that were developed in the previous longitudinal studies, particularly in NELS:88, were applied in ECLS-K, with significant improvements. The primary example of this application was the introduction in ECLS-K of real-time, two-stage adaptive testing. That is, the cognitive tests in reading, mathematics, and general knowledge were individually administered in ECLS in Grades K–1. In each subject, the score on a short routing test determined the selection of an easier or more difficult second stage form. The reading and mathematics tests each had three second-stage forms of different difficulty; two forms were used for the general knowledge test. The same assessment package was used for the first four ECLS-K rounds, fall and spring kindergarten and fall and spring first grade. The reading and mathematics test forms were designed so that, in fall kindergarten, about 75% of the sample would be expected to be routed to the easiest of the three alternate forms; by spring of first grade, the intention was that about 75% of children would receive the hardest form. Assessments for the subsequent rounds were used in only one grade. The third- and fifth-grade tests were designed to route the middle half of the sample to the middle form, with the rest receiving the easiest or most difficult form. In the eighth grade, there were only two-second stage forms, each designed to be administered to half the sample. For the routing test, each item response was entered into a portable computer by the assessor. The computer would then score the routing test responses and based on the score select the appropriate second stage form to be administered.

As in NELS:88, multiple hierarchical proficiency levels were developed to mark critical developmental points along a child's learning curve in reading and mathematics. This development was easier to do in the early rounds of ECLS-K because of the relative standardization of the curriculum in the early grades along with the generally accepted pedagogical sequencing that was followed in early mathematics and reading. When the educational treatment follows a fairly standard pedagogical sequence (as in the early grades in school), we arguably have a situation that can be characterized by a common growth curve with children located at different points along that curve signifying different levels of development. Assuming a common growth curve, the job of the test developer and the psychometrician is to identify critical points along the growth curve that mark developmental milestones. Marking these points is the task of the proficiency levels.

10.4.1 Proficiency Levels and Scores in ECLS-K

Proficiency levels as defined in ECLS-K, as in NELS:88, provide a means for distinguishing status or gain in specific skills within a content area from the overall achievement measured by the IRT scale scores. Once again, clusters of four assessment questions having similar content and difficulty were located at several points along the score scale of the reading and mathematics assessments. Each cluster marked a learning milestone in reading or mathematics, agreed on by ECLS-K curriculum specialists. The sets of proficiency levels formed a hierarchical structure in the Piagetian sense in that the teaching sequence implied that one had to master the lower levels in the sequence before one could learn the material at the next higher level. This was the same basic procedure that was introduced in NELS:88.

Clusters of four items marking critical points on the vertical score scale provide a more reliable assessment of a particular proficiency level than do single items because of the possibility of guessing. It is very unlikely that a student who has not mastered a particular skill would be able to guess enough answers correctly to pass a four-item cluster. The proficiency levels were assumed to follow a Guttman model (Guttman 1950), that is, a student passing a particular skill level was expected to have mastered all lower levels; a failure at a given level should be consistent with nonmastery at higher levels. Only a very small percentage of students in ECLS-K had response patterns that did not follow the Guttman scaling model; that is, a failing score at a lower level followed by a pass on a more difficult item cluster. (For the first five rounds of data collection, fewer than 7% of reading response patterns and fewer than 5% of mathematics assessment results failed to follow the expected hierarchical pattern.) Divergent response patterns do not necessarily indicate a different learning sequence for these children. Because all of the proficiency level items were multiple choice, a number of these reversals simply may be due to children guessing as well as other random response errors.

Sections 4.2.2 and 4.3.2 of Najarian et al. (2009) described the ten reading and nine mathematics proficiency levels identified in the kindergarten through eighth-grade assessments. No proficiency scores were computed for the science assessment because the questions did not follow a hierarchical pattern. Two types of scores were reported with respect to the proficiency levels: a single indicator of highest level mastered, and a set of IRT-based probability scores, one for each proficiency level.

10.4.2 Highest Proficiency Level Mastered

As described above, mastery of a proficiency level was defined as answering correctly at least three of the four questions in a cluster. This definition results in a very low probability of guessing enough right answers to pass a cluster by chance. The probability varies depending on the guessing parameters (IRT *c* parameters) of the items in each cluster, but is generally less than 2%. At least two incorrect or “I don’t know” responses indicated lack of mastery. Open-ended questions that were answered with an explicit “I don’t know” response were treated as wrong, while omitted items were not counted. Since the ECLS-K direct cognitive child assessment was a two-stage design (where not all children were administered all items), and since more advanced assessment instruments were administered in third grade and beyond, children’s data did not include all of the assessment items necessary to determine pass or fail for every proficiency level at each round of data collection. The missing information was not missing at random; it depended in part on children being routed to second-stage forms of varying difficulty within each assessment set and in part on different assessments being used for the different grades. In order to avoid bias due to the nonrandomness of the missing proficiency level scores, imputation procedures were undertaken to fill in the missing information.

Pass or fail for each proficiency level was based on actual counts of correct or incorrect responses, if they were present. If too few items were administered or answered to determine mastery of a level, a pass/fail score was imputed based on the remaining proficiency level scores only if they indicated a pattern that was unambiguous. That is, a fail might be inferred for a missing level if there were easier cluster(s) that had been failed and no higher cluster passed; or a pass might be assumed if harder cluster(s) were passed and no easier one failed. In the case of ambiguous patterns (e.g., pass, missing, fail for three consecutive levels, where the missing level could legitimately be either a pass or a fail), an additional imputation step was undertaken that relied on information from the child’s performance in that round of data collection on all of the items answered within the domain that included the incomplete cluster. IRT-based estimates of the probability of a correct answer were computed for each missing assessment item and used to assign an imputed right or wrong score to the item. These imputed responses were then aggregated in the same manner as actual responses to determine mastery at each of the missing levels. Over all rounds of the study, the highest level scores were determined on the basis of item response data alone for about two-thirds of reading scores and 80% for mathematics; the rest utilized IRT-based probabilities for some or all of the missing items.

The need for imputation was greatest in the eighth-grade tests, as a result of the necessary placement of the proficiency level items on either the low or high second-stage form, based on their estimated difficulty levels. Scores were not imputed for missing levels for patterns that included a reversal (e.g., fail, blank, pass) because no resolution of the missing data could result in a consistent hierarchical pattern.

Scores in the public use data file represent the highest level of proficiency mastered by each child at each round of data collection, whether this determination was made by actual item responses, by imputation, or by a combination of methods. The highest proficiency level mastered implies that children demonstrated mastery of all lower levels and nonmastery of all higher levels. A zero score indicates nonmastery of the lowest proficiency level. Scores were excluded only if the actual or imputed mastery level data resulted in a reversal pattern as defined above. The highest proficiency level-mastered scores do not necessarily correspond to an interval scale, so in analyzing the data, they should be treated as ordinal.

10.4.3 Proficiency Probability Scores and Locus of Maximum Level of Learning Gains

Proficiency probability scores are reported for each of the proficiency levels described above, at each round of data collection. With respect to their use, these scores are essentially identical to those defined in NELS:88 above. They estimate the probability of mastery of each level and can take on any value from 0 to 1. As in NELS:88, the IRT model was employed to calculate the proficiency probability scores, which indicate the probability that a child would have passed a proficiency level, based on the child's whole set of item responses in the content domain. The item clusters were treated as single items for the purpose of IRT calibration, in order to estimate students' probabilities of mastery of each set of skills. The hierarchical nature of the skill sets justified the use of the IRT model in this way.

The proficiency probability scores can be averaged to produce estimates of mastery rates within population subgroups. These continuous measures can provide an accurate look at individuals' status and change over time. Gains in probability of mastery at each proficiency level allow researchers to study not only the amount of gain in total scale score points, but also where along the score scale different children are making their largest gains in achievement during a particular time interval. That is, when a child's difference in probabilities of mastery at each of the levels computed between adjacent testing sessions is largest, say at Level 3, we can then say the child's locus of maximum level of learning gains is in the skills defined at Level 3. Locus of maximum level of learning gains is not the same thing as highest proficiency level mastered. The latter score refers to the highest proficiency level in which the child got three out of four items correct. The locus of maximum level of learning gains could well be at the next higher proficiency level. At any rate, a student's school experiences at selected times can be related to improvements in specific skills. Additional details on the use of proficiency probabilities in ECLS-K can be found in Rock and Pollack (2002a) and Rock (2007a, b).

10.5 Conclusion

One might legitimately ask: What has been the impact of the above longitudinal studies on educational policy and research? Potential influences on policy were made possible by the implementation of extensive school, teacher, parent, and student process questionnaires and their relationships with student gains. While it is difficult to pinpoint specific impacts on policy, there is considerable evidence of the usefulness of the longitudinal databases for carrying out research on policy relevant questions. For example, NCES lists more than 1,000 publications and dissertations using the NELS:88 database. Similarly, the more recent ECLS-K study lists more than 350 publications and dissertations. As already noted, the availability of a wealth of process information gathered within a longitudinal framework is a useful first step in identifying potential causal relationships between educational processes and student performance.

In summary, the main innovations that were developed primarily in NELS:88 and improved upon in ECLS-K have become standard practices in the succeeding large-scale longitudinal studies initiated by NCES. These innovations are:

- *Real-time multistage adaptive testing* to match item difficulty to each student's ability level. Such matching of item difficulty and ability reduces testing time, as well as floor and ceiling effects, while improving accuracy of measurement.
- *The implementation of multiple-group Bayesian marginal maximum likelihood procedures for item parameter and EAP score estimation*. These procedures allow the estimation of item parameters that fit both within and across longitudinal data waves. In addition, the incorporation of ability priors for subpopulations defined by the adaptive testing procedure helps in minimizing floor and ceiling effects.
- *The pooling of succeeding longitudinal data waves to re-estimate item parameters and scores*. While this full-information approach has political drawbacks since it remakes history and is somewhat inconvenient for researchers, it helps to maintain the integrity of the vertical scale and yields more accurate estimates of the score variances associated with each wave.
- *The introduction of multiple proficiency levels that mark learning milestones in a child's development*. The concept of marking a scale with multiple proficiency points is not new, but their use within the IRT model to locate where an individual is making his/her maximum gains (locus of maximum level of learning gains) is a new contribution to measuring gains. Now the longitudinal data user has three pieces of information: how much each child gains; at what skill levels he/she is making those gains; and the highest level at which he/she has demonstrated mastery.
- The concept of *relating specific gains in proficiency levels to those process variables that can be logically expected to impact changes in the skill levels marked by these proficiency levels*.

References

Adkins-Burnett, S., Meisels, S. J., & Correnti, R. (2000). Analysis to develop the third grade indirect cognitive assessments and socioemotional measures. In *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K) spring 2000 field test report*. Rockville: Westat.

Bock, D., & Aiken, M. (1981). Marginal maximum likelihood estimation of item parameters, an application of an EM algorithm. *Psychometrika, 46*, 443–459. http://dx.doi.org/10.1002/j.2333-8504.1977.tb01147.x

Braun, H. (2006). *Using the value added modeling to evaluate teaching* (Policy Information Perspective). Princeton: Educational Testing Service.

Braun, H., & Bridgeman, B. (2005). *An introduction to the measurement of change problem* (Research Memorandum No. RM-05-01). Princeton: Educational Testing Service.

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

Coleman, J. S. (1969). *Equality and achievement in education*. Boulder: Westview Press.

Coleman, J. S., & Hoffer, T. B. (1987). *Public and private schools: The impact of communities*. New York: Basic Books.

Cronbach, L. J., & Furby, L. (1970). How should we measure change—Or should we? *Psychological Bulletin, 74*, 68–80. https://doi.org/10.1037/h0029382

Cronbach, L., & Snow, R. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Ekstrom, R. B., French, J. W., & Harman, H. H. (with Dermen, D.). (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton: Educational Testing Service.

Frankel, M. R., Kohnke, L., Buonanua, D., & Tourangeau, R. (1981). *HS&B base year sample design report*. Chicago: National Opinion Research Center.

French, J. W. (1964). *Experimental comparative prediction batteries: High school and college level*. Princeton: Educational Testing Service.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Studies in social psychology in world war II* (Vol. 4). Princeton: Princeton University Press.

Heyns, B., & Hilton, T. L. (1982). The cognitive tests for high school and beyond: An assessment. *Sociology of Education, 55*, 89–102. https://doi.org/10.2307/2112290

Ingels, S. J., Scott, L. A., Rock, D. A., Pollack, J. M., & Rasinski, K. A. (1993). *NELS-88 first follow-up final technical report*. Chicago: National Opinion Research Center.

Joreskog, K., & Sorbom, D. (1996). LISREL-8: Users reference guide [Computer software manual]. Chicago: Scientific Software.

Konstantopoulos, S. (2006). Trends of school effects on student achievement: Evidence from NLS:72, HSB:82, and NELS:92. *Teachers College Record, 108*, 2550–2581. https://doi.org/10.1111/j.1467-9620.2006.00796.x

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227–242. https://doi.org/10.1007/BF02297844

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1990). BILOG-3; Item analysis and test scoring with binary logistic models [Computer software]. Chicago: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154. https://doi.org/10.2307/1165166

Muraki, E. J., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer software]. Chicago: Scientific Software.

Najarian, M., Pollack, J. M., & Sorongon, A. G., (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K), Psychometric report for the eighth grade* (NCES Report No. 2009-002). Washington, DC: National Center for Education Statistics.

National Center for Education Statistics. (2011). National longitudinal study of 1972: Overview. Retrieved from http://nces.ed.gov/surveys/nls72/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks: Sage.

Riccobono, J., Henderson, L., Burkheimer, G., Place, C., & Levensohn, J. (1981). *National longitudinal study: Data file users manual*. Washington, DC: National Center for Education Statistics.

Rock, D. A. (2007a). *A note on gain scores and their interpretation in developmental models designed to measure change in the early school years* (Research Report No. RR-07-08). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02050.x

Rock, D. A. (2007b). *Growth in reading performance during the first four years in school* (Research Report No. RR-07-39). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02081.x

Rock, D. A., & Pollack, J. M. (1991). *The NELS-88 test battery*. Washington, DC: National Center for Education Statistics.

Rock, D. A., & Pollack, J. M. (2002a). *A model based approach to measuring cognitive growth in pre-reading and reading skills during the kindergarten year* (Research Report No. RR-02-18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2002.tb01885.x

Rock, D. A., & Pollack, J. M. (2002b). *Early childhood longitudinal study–Kindergarten class of 1989–99 (ECLS-K), Psychometric report for kindergarten through the first grade* (Working Paper No. 2002–05). Washington, DC: National Center for Education Statistics.

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. 85-217). Washington, DC: National Center for Education Statistics.

Rock, D. A., Owings, J., & Lee, R. (1993). *Changes in math proficiency between 8th and 10th grades. Statistics in brief* (NCES Report No. 93-455). Washington, DC: National Center for Education Statistics.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report of the NELS: 88 base year through second follow-up* (NCES Report No. 95-382). Washington, DC: National Center for Education Statistics.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–46). Hillsdale: Erlbaum.

Scott, L. A., Rock, D. A., Pollack, J.M., & Ingels, S. J. (1995). *Two years later: Cognitive gains and school transitions of NELS: 88 eight graders. National education longitudinal study of 1998, statistical analysis report* (NCES Report No. 95-436). Washington, DC: National Center for Education Statistics.

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

Chapter 11
Evaluating Educational Programs

Samuel Ball

11.1 An Emerging Profession

Evaluating educational programs is an emerging profession, and Educational Testing Service (ETS) has played an active role in its development. The term *program evaluation* only came into wide use in the mid-1960s, when efforts at systematically assessing programs multiplied. The purpose of this kind of evaluation is to provide information to decision makers who have responsibility for existing or proposed educational programs. For instance, program evaluation may be used to help make decisions concerning whether to develop a program (*needs assessment*), how best to develop a program (*formative evaluation*), and whether to modify—or even continue—an existing program (*summative evaluation*).

Needs assessment is the process by which one identifies needs and decides upon priorities among them. *Formative evaluation* refers to the process involved when the evaluator helps the program developer—by pretesting program materials, for example. *Summative evaluation* is the evaluation of the program after it is in operation. Arguments are rife among program evaluators about what kinds of information should be provided in each of these forms of evaluation.

This chapter was written by Samuel Ball and originally published in 1979 by Educational Testing Service and later posthumously in 2011 as a research report in the ETS R&D Scientific and Policy Contributions Series. Ball was one of ETS's most active program evaluators for 10 years and directed several pacesetting studies including a large-scale evaluation of Sesame Street. The chapter documents the vigorous program of evaluation research conducted at ETS in the 1960s and 1970s, which helped lay the foundation for what was then a fledgling field. This work developed new viewpoints, techniques, and skills for systematically assessing educational programs and led to the creation of principles for program evaluation that still appear relevant today.

S. Ball (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: researchreports@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_11

In general, the ETS posture has been to try to obtain the best—that is, the most relevant, valid, and reliable—information that can be obtained within the constraints of cost and time and the needs of the various audiences for the evaluation. Sometimes, this means a tight experimental design with a national sample; at other times, the best information might be obtained through an intensive case study of a single institution. ETS has carried out both traditional and innovative evaluations of both traditional and innovative programs, and staff members also have cooperated with other institutions in planning or executing some aspects of evaluation studies. Along the way, the work by ETS has helped to develop new viewpoints, techniques, and skills.

11.2 The Range of ETS Program Evaluation Activities

Program evaluation calls for a wide range of skills, and evaluators come from a variety of disciplines: educational psychology, developmental psychology, psychometrics, sociology, statistics, anthropology, educational administration, and a host of subject matter areas. As program evaluation began to emerge as a professional concern, ETS changed, both structurally and functionally, to accommodate it. The structural changes were not exclusively tuned to the needs of conducting program evaluations. Rather, program evaluation, like the teaching of English in a well-run high school, became to some degree the concern of virtually all the professional staff. Thus, new research groups were added, and they augmented the organization's capability to conduct program evaluations.

The functional response was many-faceted. Two of the earliest evaluation studies conducted by ETS indicate the breadth of the range of interest. In 1965, collaborating with the Pennsylvania State Department of Education, Henry Dyer of ETS set out to establish a set of educational goals against which later the performance of the state's educational system could be evaluated (Dyer 1965a, b). A unique aspect of this endeavor was Dyer's insistence that the goal-setting process be opened up to strong participation by the state's citizens and not left solely to a professional or political elite. (In fact, ETS program evaluation has been marked by a strong emphasis, when at all appropriate, on obtaining community participation.)

The other early evaluation study in which ETS was involved was the now famous Coleman report (*Equality of Educational Opportunity*), issued in 1966 (Coleman et al. 1966). ETS staff, under the direction of Albert E. Beaton, had major responsibility for analysis of the massive data generated (see Beaton and Barone, Chap. 8, this volume). Until then, studies of the effectiveness of the nation's schools, especially with respect to programs' educational impact on minorities, had been small-scale. So the collection and analysis of data concerning tens of thousands of students and hundreds of schools and their communities were new experiences for ETS and for the profession of program evaluation.

In the intervening years, the Coleman report (Coleman et al. 1966) and the Pennsylvania Goals Study (Dyer 1965a, b) have become classics of their kind, and from these two auspicious early efforts, ETS has become a center of major program evaluation. Areas of focus include computer-aided instruction, aesthetics and creativity in education, educational television, educational programs for prison inmates, reading programs, camping programs, career education, bilingual education, higher education, preschool programs, special education, and drug programs. (For brief descriptions of ETS work in these areas, as well as for studies that developed relevant measures, see the appendix.) ETS also has evaluated programs relating to year-round schooling, English as a second language, desegregation, performance contracting, women's education, busing, Title I of the Elementary and Secondary Education Act (ESEA), accountability, and basic information systems.

One piece of work that must be mentioned is the *Encyclopedia of Educational Evaluation*, edited by Anderson et al. (1975). The encyclopedia contains articles by them and 36 other members of the ETS staff. Subtitled *Concepts and Techniques for Evaluating Education and Training Programs*, it contains 141 articles in all.

11.3 ETS Contributions to Program Evaluation

Given the innovativeness of many of the programs evaluated, the newness of the profession of program evaluation, and the level of expertise of the ETS staff who have directed these studies, it is not surprising that the evaluations themselves have been marked by innovations for the profession of program evaluation. At the same time, ETS has adopted several principles relative to each aspect of program evaluation. It will be useful to examine these innovations and principles in terms of the phases that a program evaluation usually attends to—goal setting, measurement selection, implementation in the field setting, analysis, and interpretation and presentation of evidence.

11.3.1 Making Goals Explicit

It would be a pleasure to report that virtually every educational program has a well-thought-through set of goals, but it is not so. It is, therefore, necessary at times for program evaluators to help verbalize and clarify the goals of a program to ensure that they are, at least, explicit. Further, the evaluator may even be given goal development as a primary task, as in the Pennsylvania Goals Study (Dyer 1965a, b). This need was seen again in a similar program, when Robert Feldmesser (1973) helped the New Jersey State Board of Education establish goals that underwrite conceptually that state's "thorough and efficient" education program.

Work by ETS staff indicates there are four important principles with respect to program goal development and explication. The first of these principles is as follows: What program developers say their program goals are may bear only a passing resemblance to what the program in fact seems to be doing.

This principle—the occasional surrealistic quality of program goals—has been noted on a number of occasions: For example, assessment instruments developed for a program evaluation on the basis of the stated goals sometimes do not seem at all sensitive to the actual curriculum. As a result, ETS program evaluators seek, whenever possible, to cooperate with program developers to help fashion the goals statement. The evaluators also will attempt to describe the program in operation and relate that description to the stated goals, as in the case of the 1971 evaluation of the second year of *Sesame Street* for Children's Television Workshop (Bogatz and Ball 1971). This comparison is an important part of the process and represents sometimes crucial information for decision makers concerned with developing or modifying a program.

The second principle is as follows: When program evaluators work cooperatively with developers in making program goals explicit, both the program and the evaluation seem to benefit.

The original *Sesame Street* evaluation (Ball and Bogatz, 1970) exemplified the usefulness of this cooperation. At the earliest planning sessions for the program, before it had a name and before it was fully funded, the developers, aided by ETS, hammered out the program goals. Thus, ETS was able to learn at the outset what the program developers had in mind, ensuring sufficient time to provide adequately developed measurement instruments. If the evaluation team had had to wait until the program itself was developed, there would not have been sufficient time to develop the instruments; more important, the evaluators might not have had sufficient understanding of the intended goals—thereby making sensible evaluation unlikely.

The third principle is as follows: There is often a great deal of empirical research to be conducted before program goals can be specified.

Sometimes, even before goals can be established or a program developed, it is necessary, through empirical research, to indicate that there is a need for the program. An illustration is provided by the research of Ruth Ekstrom and Marlaine Lockheed (1976) into the competencies gained by women through volunteer work and homemaking. The ETS researchers argued that it is desirable for women to resume their education if they wish to after years of absence. But what competencies have they picked up in the interim that might be worthy of academic credit? By identifying, surveying, and interviewing women who wished to return to formal education, Ekstrom and Lockheed established that many women had indeed learned valuable skills and knowledge. Colleges were alerted and some have begun to give credit where credit is due.

Similarly, when the federal government decided to make a concerted attack on the reading problem as it affects the total population, one area of concern was adult reading. But there was little knowledge about it. Was there an adult literacy problem? Could adults read with sufficient understanding such items as newspaper employment advertisements, shopping and movie advertisements, and bus schedules? And in investigating adult literacy, what characterized the reading tasks that should be taken into account? Murphy, in a 1973 study (Murphy 1973a), considered these factors: the *importance* of a task (the need to be able to read the material if only once a year as with income tax forms and instructions), the *intensity* of the task (a person who wants to work in the shipping department will have to read the shipping schedule each day), or the *extensivity* of the task (70% of the adult population read a newspaper but it can usually be ignored without gross problems arising). Murphy and other ETS researchers conducted surveys of reading habits and abilities, and this assessment of needs provided the government with information needed to decide on goals and develop appropriate programs.

Still a different kind of needs assessment was conducted by ETS researchers with respect to a school for learning disabled students in 1976 (Ball and Goldman 1976). The school catered to children aged 5–18 and had four separate programs and sites. ETS first served as a catalyst, helping the school's staff develop a listing of problems. Then ETS acted as an *amicus curiae*, drawing attention to those problems, making explicit and public what might have been unsaid for want of an appropriate forum. Solving these problems was the purpose of stating new institutional goals—goals that might never have been formally recognized if ETS had not worked with the school to make its needs explicit.

The fourth principle is as follows: The program evaluator should be conscious of and interested in the unintended outcomes of programs as well as the intended outcomes specified in the program's goal statement.

In program evaluation, the importance of looking for side effects, especially negative ones, has to be considered against the need to put a major effort into assessing progress toward intended outcomes. Often, in this phase of evaluation, the varying interests of evaluators, developers, and funders intersect—and professional, financial, and political considerations are all at odds. At such times, program evaluation becomes as much an art form as an exercise in social science.

A number of articles were written about this problem by Samuel J. Messick, ETS vice president for research (e.g., Messick 1970, 1975). His viewpoint—the importance of the medical model—has been illustrated in various ETS evaluation studies. His major thesis was that the medical model of program evaluation explicitly recognizes that "…prescriptions for treatment and the evaluation of their effectiveness should take into account not only reported symptoms but other characteristics of the organism and its ecology as well" (Messick 1975, p. 245). As Messick went on to point out, this characterization was a call for a systems analysis approach to program evaluation—dealing empirically with the interrelatedness of all the factors and monitoring all outcomes, not just the intended ones.

When, for example, ETS evaluated the first 2 years of *Sesame Street* (Ball and Bogatz 1970), there was obviously pressure to ascertain whether the intended goals of that show were being attained. It was nonetheless possible to look for some of the more likely unintended outcomes: whether the show had negative effects on heavy viewers going off to kindergarten, and whether the show was achieving impacts in attitudinal areas.

In summative evaluations, to study unintended outcomes is bound to cost more money than to ignore them. It is often difficult to secure increased funding for this purpose. For educational programs with potential national applications, however, ETS strongly supports this more comprehensive approach.

11.3.2 Measuring Program Impact

The letters *ETS* have become almost synonymous in some circles with standardized testing of student achievement. In its program evaluations, ETS naturally uses such tests as appropriate, but frequently the standardized tests are not appropriate measures. In some evaluations, ETS uses both standardized and domain-referenced tests. An example may be seen in *The Electric Company* evaluations (Ball et al. 1974). This televised series, which was intended to teach reading skills to first through fourth graders, was evaluated in some 600 classrooms. One question that was asked during the process concerned the interaction of the student's level of reading attainment and the effectiveness of viewing the series. Do good readers learn more from the series than poor readers? So standardized, norm-referenced reading tests were administered, and the students in each grade were divided into deciles on this basis, thereby yielding ten levels of reading attainment.

Data on the outcomes using the domain-referenced tests were subsequently analyzed for each decile ranking. Thus, ETS was able to specify for what level of reading attainment, in each grade, the series was working best. This kind of conclusion would not have been possible if a specially designed domain-referenced reading test with no external referent had been the only one used, nor if a standardized test, not sensitive to the program's impact, had been the only one used.

Without denying the usefulness of previously designed and developed measures, ETS evaluators have frequently preferred to develop or adapt instruments that would be specifically sensitive to the tasks at hand. Sometimes this measurement effort is carried out in anticipation of the needs of program evaluators for a particular instrument, and sometimes because a current program evaluation requires immediate instrumentation.

An example of the former is a study of doctoral programs by Mary Jo Clark et al. (1976). Existing instruments had been based on surveys in which practitioners in a given discipline were asked to rate the quality of doctoral programs in that discipline. Instead of this reputational survey approach, the ETS team developed an array of criteria (e.g., faculty quality, student body quality, resources, academic offerings, alumni performance), all open to objective assessment. This assessment tool can be used to assess changes in the quality of the doctoral programs offered by major universities.

Similarly, the development by ETS of the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom et al. 1976) also provided a tool—one that could be used when evaluating the cognitive abilities of teachers or students if these structures were of interest in a particular evaluation. A clearly useful application was in the California study of teaching performance by Frederick McDonald and Patricia Elias (1976). Teachers with certain kinds of cognitive structures were seen to have differential impacts on student achievement. In the Donald A. Trismen study of an aesthetics program (Trismen 1968), the factor kit was used to see whether cognitive structures interacted with aesthetic judgments.

11.3.2.1 Developing Special Instruments

Examples of the development of specific instrumentation for ETS program evaluations are numerous. Virtually every program evaluation involves, at the very least, some adapting of existing instruments. For example, a questionnaire or interview may be adapted from ones developed for earlier studies. Typically, however, new instruments, including goal-specific tests, are prepared. Some ingenious examples, based on the 1966 work of E. J. Webb, D. F. Campbell, R. D. Schwartz, and L. Sechrest, were suggested by Anderson (1968) for evaluating museum programs, and the title of her article gives a flavor of the unobtrusive measures illustrated—“Noseprints on the Glass.”

Another example of ingenuity is Trismen’s use of 35 mm slides as stimuli in the assessment battery of the Education through Vision program (Trismen 1968). Each slide presented an art masterpiece, and the response options were four abstract designs varying in color. The instruction to the student was to pick the design that best illustrated the masterpiece’s coloring.

11.3.2.2 Using Multiple Measures

When ETS evaluators have to assess a variable and the usual measures have rather high levels of error inherent in them, they usually resort to triangulation. That is, they use multiple measures of the same construct, knowing that each measure suffers from a specific weakness. Thus, in 1975, Donald E. Powers evaluated for the Philadelphia school system the impact of dual-audio television—a television show telecast at the same time as a designated FM radio station provided an appropriate educational commentary. One problem in measurement was assessing the amount of contact the student had with the dual-audio television treatment (Powers 1975a). Powers used home telephone interviews, student questionnaires, and very simple knowledge tests of the characters in the shows to assess whether students had in fact been exposed to the treatment. Each of these three measures has problems associated with it, but the combination provided a useful assessment index.

In some circumstances, ETS evaluators are able to develop measurement techniques that are an integral part of the treatment itself. This unobtrusiveness has clear benefits and is most readily attainable with computer-aided instructional (CAI) programs. Thus, for example, Donald L. Alderman, in the evaluation of TICCIT (a CAI program developed by the Mitre Corporation), obtained for each student such indices as the number of lessons passed, the time spent on line, the number of errors made, and the kinds of errors (Alderman 1978). And he did this simply by programming the computer to save this information over given periods of time.

11.3.3 Working in Field Settings

Measurement problems cannot be addressed satisfactorily if the setting in which the measures are to be administered is ignored. One of the clear lessons learned in ETS program evaluation studies is that measurement in field settings (home, school, community) poses different problems from measurement conducted in a laboratory.

Program evaluation, ether formative or summative, demands that its empirical elements usually be conducted in natural field settings rather than in more contrived settings, such as a laboratory. Nonetheless, the problems of working in field settings are rarely systematically discussed or researched. In an article in the *Encyclopedia of Educational Evaluation*, Bogatz (1975) detailed these major aspects:

- Obtaining permission to collect data at a site
- Selecting a field staff
- Training the staff
- Maintaining family/community support

Of course, all the aspects discussed by Bogatz interact with the measurement and design of the program evaluation. A great source of information concerning field operations is the ETS Head Start Longitudinal Study of Disadvantaged Children, directed by Virginia Shipman (1970). Although not primarily a program evaluation, it certainly has generated implications for early childhood programs. It was longitudinal, comprehensive in scope, and large in size, encompassing four sites and, initially, some 2000 preschoolers. It was clear from the outset that close community ties were essential if only for expediency—although, of course, more important ethical principles were involved. This close relationship with the communities in which the study was conducted involved using local residents as supervisors and testers, establishing local advisory committees, and thus ensuring free, two-way communication between the research team and the community.

The *Sesame Street* evaluation also adopted this approach (Ball and Bogatz 1970). In part because of time pressures and in part to ensure valid test results, the ETS evaluators especially developed the tests so that community members with minimal educational attainments could be trained quickly to administer them with proper skill.

11.3.3.1 Establishing Community Rapport

In evaluations of street academies by Ronald L. Flaugher (1971), and of education programs in prisons by Flaugher and Samuel Barnett (1972), it was argued that one of the most important elements in successful field relationships is the time an evaluator spends getting to know the interests and concerns of various groups, and lowering barriers of suspicion that frequently separate the educated evaluator and the less-educated program participants. This point may not seem particularly sophisticated or complex, but many program evaluations have floundered because of an evaluator's lack of regard for disadvantaged communities (Anderson 1970). Therefore, a firm principle underlying ETS program evaluation is to be concerned with the communities that provide the contexts for the programs being evaluated. Establishing two-way lines of communication with these communities and using community resources whenever possible help ensure a valid evaluation.

Even with the best possible community support, field settings cause problems for measurement. Raymond G. Wasdyke and Jerilee Grandy (1976) showed this idea to be true in an evaluation in which the field setting was literally that—a field setting. In studying the impact of a camping program on New York City grade school pupils, they recognized the need, common to most evaluations, to describe the treatment—in this case the camping experience. Therefore, ETS sent an observer to the campsite with the treatment groups. This person, who was herself skilled in camping, managed not to be an obtrusive participant by maintaining a relatively low profile.

Of course, the problems of the observer can be just as difficult in formal institutions as on the campground. In their 1974 evaluation of Open University materials, Hartnett and colleagues found, as have program evaluators in almost every situation, that there was some defensiveness in each of the institutions in which they worked (Hartnett et al. 1974). Both personal and professional contacts were used to allay suspicions. There also was emphasis on an evaluation design that took into account each institution's values. That is, part of the evaluation was specific to the institution, but some common elements across institutions were retained. This strategy underscored the evaluators' realization that each institution was different, but allowed ETS to study certain variables across all three participating institutions.

Breaking down the barriers in a field setting is one of the important elements of a successful evaluation, yet each situation demands somewhat different evaluator responses.

11.3.3.2 Involving Program Staff

Another way of ensuring that evaluation field staff are accepted by program staff is to make the program staff active participants in the evaluation process. While this integration is obviously a technique to be strongly recommended in formative evaluations, it can also be used in summative evaluations. In his evaluation of PLATO in junior colleges, Murphy (1977) could not afford to become the victim of a program developer's fear of an insensitive evaluator. He overcame this potential problem by enlisting the active participation of the junior college and program development staffs. One of Murphy's concerns was that there is no common course across colleges. Introduction to Psychology, for example, might be taught virtually everywhere, but the content can change remarkably, depending on such factors as who teaches the course, where it is taught, and what text is used. Murphy understood this variability and his evaluation of PLATO reflected his concern. It also necessitated considerable input and cooperation from program developers and college teachers working in concert—with Murphy acting as the conductor.

11.3.4 Analyzing the Data

After the principles and strategies used by program evaluators in their field operations are successful and data are obtained, there remains the important phase of data analysis. In practice, of course, the program evaluator thinks through the question of data analysis *before* entering the data collection phase. Plans for analysis help determine what measures to develop, what data to collect, and even, to some extent, how the field operation is to be conducted. Nonetheless, analysis plans drawn up early in the program evaluation cannot remain quite as immutable as the Mosaic Law. To illustrate the need for flexibility, it is useful to turn once again to the heuristic ETS evaluation of *Sesame Street*.

As initially planned, the design of the *Sesame Street* evaluation was a true experiment (Ball and Bogatz 1970). The analyses called for were multivariate analyses of covariance, using pretest scores as the covariate. At each site, a pool of eligible preschoolers was obtained by community census, and experimental and control groups were formed by random assignment from these pools. The evaluators were somewhat concerned that those designated to be the experimental (viewing) group might not view the show—it was a new show on public television, a loose network of TV stations not noted for high viewership. Some members of the *Sesame Street* national research advisory committee counseled ETS to consider paying the experimental group to view. The suggestion was resisted, however, because any efforts above mild and occasional verbal encouragement to view the show would compromise the results. If the experimental group members were paid, and if they then viewed extensively and outperformed the control group at posttest, would the improved performance be due to the viewing, the payment, or some interaction of payment and viewing? Of course, this nice argument proved to be not much more than an exercise in modern scholasticism. In fact, the problem lay not in the treatment group but in the uninformed and unencouraged-to-view control group. The members of that group, as indeed preschoolers with access to public television throughout the nation, were viewing the show with considerable frequency—and not much less than the experimental group. Thus, the planned analysis involving differences in posttest attainments between the two groups was dealt a mortal blow.

Fortunately, other analyses were available, of which the ETS-refined age cohorts design provided a rational basis. This design is presented in the relevant report (Ball and Bogatz 1970). The need here is not to describe the design and analysis but to emphasize a point made practically by the poet Robert Burns some time ago and repeated here more prosaically: The best laid plans of evaluators can “gang aft agley,” too.

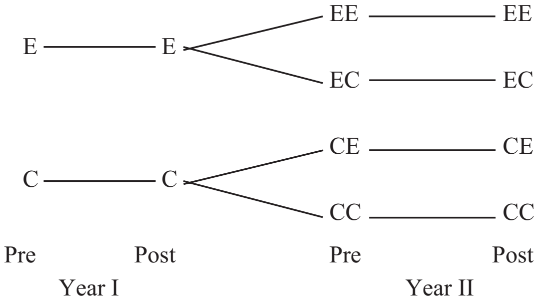


Fig. 11.1 The design for the new pool of classes. For Year II, EE represents children who were in E classrooms in Year I and again in Year II. That is, the first letter refers to status in Year I and the second to status in Year II

11.3.4.1 Clearing New Paths

Sometimes program evaluators find that the design and analysis they have in mind represent an untrodden path. This result is perhaps in part because many of the designs in the social sciences are built upon laboratory conditions and simply are not particularly relevant to what happens in educational institutions.

When ETS designed the summative evaluation of *The Electric Company*, it was able to set up a true experiment in the schools. Pairs of comparable classrooms within a school and within a grade were designated as the pool with which to work. One of each pair of classes was randomly assigned to view the series. Pretest scores were used as covariates on posttest scores, and in 1973 the first-year evaluation analysis was successfully carried out (Ball and Bogatz 1973). The evaluation was continued through a second year, however, and as is usual in schools, the classes did not remain intact.

From an initial 200 classes, the children had scattered through many more classrooms. Virtually none of the classes with subject children contained only experimental or only control children from the previous year. Donald B. Rubin, an ETS statistician, consulted with a variety of authorities and found that the design and analysis problem for the second year of the evaluation had not been addressed in previous work. To summarize the solution decided on, the new pool of classes was reassigned randomly to *E* (experimental) or *C* (control) conditions so that over the 2 years the design was portrayable as Fig. 11.1.

Further, the pretest scores of Year II were usable as new covariates when analyzing the results of the Year II posttest scores (Ball et al. 1974).

11.3.4.2 Tailoring to the Task

Unfortunately for those who prefer routine procedures, it has been shown across a wide range of ETS program evaluations that each design and analysis must be tailored to the occasion. Thus, Gary Marco (1972), as part of the statewide educational assessment in Michigan, evaluated ESEA Title I program performance. He assessed the amount of exposure students had to various clusters of Title I programs, and he included control schools in the analysis. He found that a regression-analysis model involving a correction for measurement error was an innovative approach that best fit his complex configuration of data.

Garlie Forehand, Marjorie Ragosta, and Donald A. Rock, in a national, correlational study of desegregation, obtained data on school characteristics and on student outcomes (Forehand et al. 1976). The purposes of the study included defining indicators of effective desegregation and discriminating between more and less effective school desegregation programs. The emphasis throughout the effort was on variables that were manipulable. That is, the idea was that evaluators would be able to suggest practical advice on what schools can do to achieve a productive desegregation program. Initial investigations allowed specification among the myriad variables of a hypothesized set of causal relationships, and the use of path analysis made possible estimation of the strength of hypothesized causal relationships. On the basis of the initial correlation matrices, the path analyses, and the observations made during the study, an important product—a nontechnical handbook for use in schools—was developed.

Another large-scale ETS evaluation effort was directed by Trismen et al. (1976). They studied compensatory reading programs, initially surveying more than 700 schools across the country. Over a 4-year period ending in 1976, this evaluation interspersed data analysis with new data collection efforts. One purpose was to find schools that provided exceptionally positive or negative program results. These schools were visited blind and observed by ETS staff. Whereas the Forehand evaluation analysis (Forehand et al. 1976) was geared to obtaining practical applications, the equally extensive evaluation analysis of Trismen's study was aimed at generating hypotheses to be tested in a series of smaller experiments.

As a further illustration of the complex interrelationship among evaluation purposes, design, analyses, and products, there is the 1977 evaluation of the use of PLATO in the elementary school by Spencer Swinton and Marianne Amarel (1978). They used a form of regression analysis—as did Forehand et al. (1976) and Trismen et al. (1976). But here the regression analyses were used differently in order to identify program effects unconfounded by teacher differences. In this regression analysis, teachers became fixed effects, and contrasts were fitted for each within-teacher pair (experimental versus control classroom teachers).

This design, in turn, provides a contrast to McDonald's (1977) evaluation of West New York programs to teach English as a second language to adults. In this instance, the regression analysis was directed toward showing which teaching method related most to gains in adult students' performance.

There is a school of thought within the evaluation profession that design and analysis in program evaluation can be made routine. At this point, the experience of ETS indicates that this would be unwise.

11.3.5 *Interpreting the Results*

Possibly the most important principle in program evaluation is that interpretations of the evaluation's meaning—the conclusions to be drawn—are often open to various nuances. Another problem is that the evidence on which the interpretations are based may be inconsistent. The initial premise of this chapter was that the role of program evaluation is to provide evidence for decision-makers. Thus, one could argue that differences in interpretation, and inconsistencies in the evidence, are simply problems for the decision-maker and not for the evaluator.

But consider, for example, an evaluation by Powers of a year-round program in a school district in Virginia (Powers 1974, 1975b). (The long vacation was staggered around the year so that schools remained open in the summer.) The evidence presented by Powers indicated that the year-round school program provided a better utilization of physical plant and that student performance was not negatively affected. The school board considered this evidence as well as other conflicting evidence provided by Powers that the parents' attitudes were decidedly negative. The board made up its mind, and (not surprisingly) scotched the program. Clearly, however, the decision was not up to Powers. His role was to collect the evidence and present it systematically.

11.3.5.1 Keeping the Process Open

In general, the ETS response to conflicting evidence or varieties of nuances in interpretation is to keep the evaluation process and its reporting as open as possible. In this way, the values of the evaluator, though necessarily present, are less likely to be a predominating influence on subsequent action.

Program evaluators do, at times, have the opportunity to influence decision-makers by showing them that there are kinds of evidence not typically considered. The Coleman Study, for example, showed at least some decision-makers that there is more to evaluating school programs than counting (or calculating) the numbers of books in libraries, the amount of classroom space per student, the student-teacher ratio, and the availability of audiovisual equipment (Coleman et al. 1966). Rather, the output of the schools in terms of student performance was shown to be generally superior as evidence of school program performance.

Through their work, evaluators are also able to educate decision makers to consider the important principle that educational treatments may have positive effects for some students and negative effects for others—that an interaction of treatment with student should be looked for. As pointed out in the discussion of unintended outcomes, a systems-analysis approach to program evaluation—dealing empirically with the interrelatedness of all the factors that may affect performance—is to be preferred. And this approach, as Messick emphasized, "properly takes into account those student-process-environment interactions that produce differential results" (Messick 1975, p. 246).

11.3.5.2 Selecting Appropriate Evidence

Finally, a consideration of the kinds of evidence and interpretations to be provided decision makers leads inexorably to the realization that different kinds of evidence are needed, depending on the decision-maker's problems and the availability of resources. The most scientific evidence involving objective data on student performance can be brilliantly interpreted by an evaluator, but it might also be an abomination to a decision maker who really needs to know whether teachers' attitudes are favorable.

ETS evaluations have provided a great variety of evidence. For a formative evaluation in Brevard County, Florida, Trismen (1970) provided evidence that students could make intelligent choices about courses. In the ungraded schools, students had considerable freedom of choice, but they and their counselors needed considerably more information than in traditional schools about the ingredients for success in each of the available courses. As another example, Gary Echternacht, George Temp, and Theodore Stolie helped state and local education authorities develop Title I reporting models that included evidence on impact, cost, and compliance with federal regulations (Echternacht et al. 1976). Forehand and McDonald (1972) had been working with New York City to develop an accountability model providing constructive kinds of evidence for the city's school system. On the other hand, as part of an evaluation team, Amarel provided, for a small experimental school in Chicago, judgmental data as well as reports and documents based on the school's own records and files (Amarel and The Evaluation Collective 1979). Finally, Michael Rosenfeld provided Montgomery Township, New Jersey, with student, teacher, and parent perceptions in his evaluation of the open classroom approach then being tried out (Rosenfeld 1973).

In short, just as tests are not valid or invalid (it is the ways tests are used that deserve such descriptions), so too, evidence is not good or bad until it is seen in relation to the purpose for which it is to be used, and in relation to its utility to decision-makers.

11.4 Postscript

For the most part, ETS's involvement in program evaluation has been at the practical level. Without an accompanying concern for the theoretical and professional issues, however, practical involvement would be irresponsible. ETS staff members have therefore seen the need to integrate and systematize knowledge about program evaluation. Thus, Anderson obtained a contract with the Office of Naval Research to draw together the accumulated knowledge of professionals from inside and outside ETS on the topic of program evaluation. A number of products followed. These products included a survey of practices in program evaluation (Ball and Anderson 1975a), and a codification of program evaluation principles and issues (Ball and

Anderson 1975b). Perhaps the most generally useful of the products is the aforementioned *Encyclopedia of Educational Evaluation* (Anderson et al. 1975).

From an uncoordinated, nonprescient beginning in the mid-1960s, ETS has acquired a great deal of experience in program evaluation. In one sense it remains uncoordinated because there is no specific "party line," no dogma designed to ensure ritualized responses. It remains quite possible for different program evaluators at ETS to recommend differently designed evaluations for the same burgeoning or existing programs.

There is no sure knowledge where the profession of program evaluation is going. Perhaps, with zero-based budgeting, program evaluation will experience amazing growth over the next decade, growth that will dwarf its current status (which already dwarfs its status of a decade ago). Or perhaps there will be a revulsion against the use of social scientific techniques within the political, value-dominated arena of program development and justification. At ETS, the consensus is that continued growth is the more likely event. And with the staff's variegated backgrounds and accumulating expertise, ETS hopes to continue making significant contributions to this emerging profession.

Appendix: Descriptions of ETS Evaluation and Some Related Studies in Some Key Categories

Aesthetics and Creativity in Education

For Bartlett Hayes III's program of Education through Vision at Andover Academy, Donald A. Trismen developed a battery of evaluation instruments that assessed, inter alia, a variety of aesthetic judgments (Trismen 1968). Other ETS staff members working in this area have included Norman Frederiksen and William C. Ward, who have developed a variety of assessment techniques for tapping creativity and scientific creativity (Frederiksen and Ward 1975; Ward and Frederiksen 1977); Richard T. Murphy, who also has developed creativity-assessing techniques (Murphy 1973b, 1977); and Scarvia B. Anderson, who described a variety of ways to assess the effectiveness of aesthetic displays (Anderson 1968).

Bilingual Education

ETS staff have conducted and assisted in evaluations of numerous and varied programs of bilingual education. For example, Berkeley office staff (Reginald A. Corder, Patricia Elias, Patricia Wheeler) have evaluated programs in Calexico (Corder 1976a), Hacienda-La Puente (Elias and Wheeler 1972), and El Monte (Corder and Johnson 1972). For the Los Angeles office, J. Richard Harsh (1975) evaluated a bilingual program in Azusa, and Ivor Thomas (1970) evaluated one in Fountain Valley. Donald E. Hood (1974) of the Austin office evaluated the Dallas Bilingual Multicultural Program. These evaluations were variously formative and summative and covered bilingual programs that, in combination, served students from preschool (Fountain Valley) through 12th grade (Calexico).

Camping Programs

Those in charge of a school camping program in New York City felt that it was having unusual and positive effects on the students, especially in terms of motivation. ETS was asked to—and did—evaluate this program, using an innovative design and measurement procedures developed by Raymond G. Wasdyke and Jerilee Grandy (1976).

Career Education

In a decade of heavy federal emphasis on career education, ETS was involved in the evaluation of numerous programs in that field. For instance, Raymond G. Wasdyke (1977) helped the Newark, Delaware, school system determine whether its career education goals and programs were properly meshed. In Dallas, Donald Hood (1972) of the ETS regional staff assisted in developing goal specifications and reviewing evaluation test items for the Skyline Project, a performance contract calling for the training of high school students in 12 career clusters. Norman E. Freeberg (1970) developed a test battery to be used in evaluating the Neighborhood Youth Corps. Ivor Thomas (1973) of the Los Angeles office provided formative evaluation services for the Azusa Unified School District's 10th grade career training and performance program for disadvantaged students. Roy Hardy (1977) of the Atlanta office directed the third-party evaluation of Florida's Comprehensive Program of Vocational Education for Career Development, and Wasdyke (1976) evaluated the Maryland Career Information System. Reginald A. Corder, Jr. (1975) of the Berkeley office assisted in the evaluation of the California Career Education program and subsequently directed the evaluation of the Experience-Based Career Education Models of a number of regional education laboratories (Corder 1976b).

Computer-Aided Instruction

Three major computer-aided instruction programs developed for use in schools and colleges have been evaluated by ETS. The most ambitious is PLATO from the University of Illinois. Initially, the ETS evaluation was directed by Ernest Anastasio

(1972), but later the effort was divided between Richard T. Murphy, who focused on college-level programs in PLATO, and Spencer Swinton and Marianne Amarel (1978), who focused on elementary and secondary school programs. ETS also directed the evaluation of TICCIT, an instructional program for junior colleges that used small-computer technology; the study was conducted by Donald L. Alderman (1978). Marjorie Ragosta directed the evaluation of the first major in-school longitudinal demonstration of computer-aided instruction for low-income students (Holland et al. 1976).

Drug Programs

Robert F. Boldt (1975) served as a consultant on the National Academy of Science's study assessing the effectiveness of drug antagonists (less harmful drugs that will "fight" the impact of illegal drugs). Samuel Ball (1973) served on a National Academy of Science panel that designed, for the National Institutes of Health, a means of evaluating media drug information programs and spot advertisements.

Educational Television

ETS was responsible for the national summative evaluation of the ETV series *Sesame Street* for preschoolers (Ball and Bogatz 1970), and *The Electric Company* for students in Grades 1 through 4 (Ball and Bogatz 1973); the principal evaluators were Samuel Ball, Gerry Ann Bogatz, and Donald B. Rubin. Additionally, Ronald Flaugher and Joan Knapp (1972) evaluated the series *Bread and Butterflies* to clarify career choice; Jayjia Hsia (1976) evaluated a series on the teaching of English for high school students and a series on parenting for adults.

Higher Education

Much ETS research in higher education focuses on evaluating students or teachers, rather than programs, mirroring the fact that systematic program evaluation is not common at this level. ETS has made, however, at least two major forays in program evaluation in higher education. In their Open University study, Rodney T. Hartnett and associates joined with three American universities (Houston, Maryland, and Rutgers) to see if the British Open University's methods and materials were appropriate for American institutions Hartnett et al. 1974). Mary Jo Clark, Leonard L. Baird, and Hartnett conducted a study of means of assessing quality in doctoral programs (Clark et al. 1976). They established an array of criteria for use in obtaining more precise descriptions and evaluations of doctoral programs than the prevailing technique—reputational surveys—provides. P. R. Harvey (1974) also evaluated the National College of Education Bilingual Teacher Education project, while Protase Woodford, (1975) proposed a pilot project for oral proficiency interview tests of bilingual teachers and tentative determination of language proficiency criteria.

Preschool Programs

A number of preschool programs have been evaluated by ETS staff, including the ETV series *Sesame Street* (Ball and Bogatz 1970; Bogatz and Ball 1971). Irving Sigel (1976) conducted formative studies of developmental curriculum. Virginia Shipman (1974) helped the Bell Telephone Companies evaluate their day care centers, Samuel Ball, Brent Bridgeman, and Albert Beaton provided the U.S. Office of Child Development with a sophisticated design for the evaluation of Parent-Child Development Centers (Ball et al. 1976), and Ball and Kathryn Kazarow evaluated the To Reach a Child program (Ball and Kazarow 1974). Roy Hardy (1975) examined the development of CIRCO, a Spanish language test battery for preschool children.

Prison Programs

In New Jersey, ETS has been involved in the evaluation of educational programs for prisoners. Developed and administered by Mercer County Community College, the programs have been subject to ongoing study by Ronald L. Flaugher and Samuel Barnett (1972).

Reading Programs

ETS evaluators have been involved in a variety of ways in a variety of programs and proposed programs in reading. For example, in an extensive, national evaluation, Donald A. Trismen et al. (1976) studied the effectiveness of reading instruction in compensatory programs. At the same time, Donald E. Powers (1973) conducted a small study of the impact of a local reading program in Trenton, New Jersey. Ann M. Bussis, Edward A. Chittenden, and Marianne Amarel reported the results of their study of primary school teachers' perceptions of their own teaching behavior (Bussis et al. 1976). Earlier, Richard T. Murphy surveyed the reading competencies and needs of the adult population (Murphy 1973a).

Special Education

Samuel Ball and Karla Goldman (1976) conducted an evaluation of the largest private school for the learning disabled in New York City, and Carol Vale (1975) of the ETS office in Berkeley directed a national needs assessment concerning educational technology and special education. Paul Campbell (1976) directed a major study of an intervention program for learning disabled juvenile delinquents.

References

Alderman, D. L. (1978). *Evaluation of the TICCIT computer-assisted instructional system in the community college*. Princeton: Educational Testing Service.

Amarel, M., & The Evaluation Collective. (1979). *Reform, response, renegotiation: Transitions in a school-change project.* Unpublished manuscript.

Anastasio, E. J. (1972). *Evaluation of the PLATO and TICCIT computer-based instructional systems—A preliminary plan* (Program Report No. PR-72-19). Princeton: Educational Testing Service.

Anderson, S. B. (1968). Noseprints on the glass—Or how do we evaluate museum programs? In E. Larrabee (Ed.), *Museums and education* (pp. 115–126). Washington, DC: Smithsonian Institution Press.

Anderson, S. B. (1970). From textbooks to reality: Social researchers face the facts of life in the world of the disadvantaged. In J. Hellmuth (Ed.), *Disadvantaged child: Vol. 3. Compensatory education: A national debate*. New York: Brunner/Mazel.

Anderson, S. B., Ball, S., & Murphy, R. T. (Eds.). (1975). *Encyclopedia of educational evaluation: Concepts and techniques for evaluating education and training programs*. San Francisco: Jossey-Bass Publishers.

Ball, S. (1973, July). *Evaluation of drug information programs—Report of the panel on the impact of information on drug use and misuse, phase 2*. Washington, DC: National Research Council, National Academy of Sciences.

Ball, S., & Anderson, S. B. (1975a). *Practices in program evaluation: A survey and some case studies*. Princeton: Educational Testing Service.

Ball, S., & Anderson, S. B. (1975b). *Professional issues in the evaluation of education/training programs*. Princeton: Educational Testing Service.

Ball, S., & Bogatz, G. A. (1970). *The first year of Sesame Street: An evaluation* (Program Report No. PR-70-15). Princeton: Educational Testing Service.

Ball, S., & Bogatz, G. A. (1973). *Reading with television: An evaluation of the Electric Company* (Program Report No. PR-73-02). Princeton: Educational Testing Service.

Ball, S., & Goldman, K. S. (1976). *The Adams School An interim report*. Princeton: Educational Testing Service.

Ball, S., & Kazarow, K. M. (1974). *Evaluation of To Reach a Child*. Princeton: Educational Testing Service.

Ball, S., Bogatz, G. A., Kazarow, K. M., & Rubin, D. B. (1974). *Reading with television: A follow-up evaluation of The Electric Company* (Program Report No. PR-74-15). Princeton: Educational Testing Service.

Ball, S., Bridgeman, B., & Beaton, A. E. (1976). *A design for the evaluation of the parent-child development center replication project*. Princeton: Educational Testing Service.

Bogatz, G. A. (1975). Field operations. In S. B. Anderson, S. Ball, & R. T. Murphy (Eds.), *Encyclopedia of educational evaluation* (pp. 169–175). San Francisco: Jossey-Bass Publishers.

Bogatz, G. A., & Ball, S. (1971). *The second year of Sesame Street: A continuing evaluation* (Program Report No. PR-71-21). Princeton: Educational Testing Service.

Boldt, R. F. (with Gitomer, N.). (1975). *Editing and scaling of instrument packets for the clinical evaluation of narcotic antagonists* (Program Report No. PR-75-12). Princeton: Educational Testing Service.

Bussis, A. M., Chittenden, E. A., & Amarel, M. (1976). *Beyond surface curriculum. An interview study of teachers' understandings*. Boulder: Westview Press.

Campbell, P. B. (1976). *Psychoeducational diagnostic services for learning disabled youths* [Proposal submitted to Creighton Institute for Business Law and Social Research]. Princeton: Educational Testing Service.

Clark, M. J., Hartnett, R. Y., & Baird, L. L. (1976). *Assessing dimensions of quality in doctoral education* (Program Report No. PR-76-27*)*. Princeton: Educational Testing Service.

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.

Corder, R. A. (1975). *Final evaluation report of part C of the California career education program*. Berkeley: Educational Testing Service.

Corder, R. A. (1976a). *Calexico intercultural design. El Cid Title VII yearly final evaluation reports for grades 7–12 of program of bilingual education, 1970–1976*. Berkeley: Educational Testing Service.

Corder, R. A. (1976b). *External evaluator's final report on the experience-based career education program*. Berkeley: Educational Testing Service.

Corder, R. A., & Johnson, S. (1972). *Final evaluation report, 1971–1972, MANO A MANO*. Berkeley: Educational Testing Service.

Dyer, H. S. (1965a). *A plan for evaluating the quality of educational programs in Pennsylvania* (Vol. 1, pp 1–4, 10–12). Harrisburg: State Board of Education.

Dyer, H. S. (1965b). *A plan for evaluating the quality of educational programs in Pennsylvania* (Vol. 2, pp. 158–161). Harrisburg: State Board of Education.

Echternacht, G., Temp, G., & Storlie, T. (1976). *The operation of an ESEA Title I evaluation technical assistance center—Region* 2 [Proposal submitted to DHEW/O]. Princeton: Educational Testing Service.

Ekstrom, R. B., & Lockheed, M. (1976). Giving women college credit where credit is due. *Findings, 3*(3), 1–5.

Ekstrom, R. B., French, J., & Harman, H. (with Dermen, D.). (1976). *Kit of factor-referenced cognitive tests*. Princeton: Educational Testing Service.

Elias, P., & Wheeler, P. (1972). *Interim evaluation report: BUENO*. Berkeley: Educational Testing Service.

Feldmesser, R. A. (1973). *Educational goal indicators for New Jersey* (Program Report No. PR-73-01). Princeton: Educational Testing Service.

Flaugher, R. L. (1971). *Progress report on the activities of ETS for the postal academy program.* Unpublished manuscript, Educational Testing Service, Princeton.

Flaugher, R., & Barnett, S. (1972). *An evaluation of the prison educational network*. Unpublished manuscript, Educational Testing Service, Princeton.

Flaugher, R., & Knapp, J. (1972). *Report on evaluation activities of the Bread and Butterflies project*. Princeton: Educational Testing Service.

Forehand, G. A., & McDonald, F. J. (1972). *A design for an accountability system for the New York City school system*. Princeton: Educational Testing Service.

Forehand, G. A., Ragosta, M., & Rock, D. A. (1976). *Final report: Conditions and processes of effective school desegregation* (Program Report No. PR-76-23). Princeton: Educational Testing Service.

Frederiksen, N., & Ward, W. C. (1975). *Development of measures for the study of creativity* (Research Bulletin No. RB-75-18). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01058.x

Freeberg, N. E. (1970). Assessment of disadvantaged adolescents: A different approach to research and evaluation measures. *Journal of Educational Psychology, 61*, 229–240. https://doi.org/10.1037/h0029243

Hardy, R. A. (1975). *CIRCO: The development of a Spanish language test battery for preschool children.* Paper presented at the Florida Educational Research Association, Tampa, FL.

Hardy, R. (1977). *Evaluation strategy for developmental projects in career education*. Tallahassee: Florida Department of Education, Division of Vocational, Technical, and Adult Education.

Harsh, J. R. (1975). *A bilingual/bicultural project. Azusa unified school district evaluation summary*. Los Angeles: Educational Testing Service.

Hartnett, R. T., Clark, M. J., Feldmesser, R. A., Gieber, M. L., & Soss, N. M. (1974). *The British Open University in the United States*. Princeton: Educational Testing Service.

Harvey, P. R. (1974). *National College of Education bilingual teacher education project*. Evanston: Educational Testing Service.

Holland, P. W., Jamison, D. T., & Ragosta, M. (1976). *Project report no. 1—Phase 1 final report research design*. Princeton: Educational Testing Service.

Hood, D. E. (1972). *Final audit report: Skyline career development center*. Austin: Educational Testing Service.

Hood, D. E. (1974). *Final audit report of the ESEA IV supplementary reading programs of the Dallas Independent School District. Bilingual education program*. Austin: Educational Testing Service.

Hsia, J. (1976). *Proposed formative evaluation of a WNET/*13 *pilot television program: The Speech Class* [Proposal submitted to educational broadcasting corporation]. Princeton: Educational Testing Service.

Marco, G. L. (1972). *Impact of Michigan 1970–71 grade 3 title I reading programs* (Program Report No. PR-72-05). Princeton: Educational Testing Service.

McDonald, F. J. (1977). *The effects of classroom interaction patterns and student characteristics on the acquisition of proficiency in English as a second language* (Program Report No. PR-77-05). Princeton: Educational Testing Service.

McDonald, F. J., & Elias, P. (1976). *Beginning teacher evaluation study, Phase 2. The effects of teaching performance on pupil learning* (Vol. 1, Program Report No. PR-76-06A). Princeton: Educational Testing Service.

Messick, S. (1970). The criterion problem in the evaluation of instruction: Assessing possible, not just intended outcomes. In M. Wittrock & D. Wiley (Eds.), *The evaluation of instruction: Issues and problems* (pp. 183–220). New York: Holt, Rinehart and Winston.

Messick, S. (1975). Medical model of evaluation. In S. B. Anderson, S. Ball, & R. T. Murphy (Eds.), *Encyclopedia of educational evaluation* (pp. 245–247). San Francisco: Jossey-Bass Publishers.

Murphy, R. T. (1973a). *Adult functional reading study* (Program Report No. PR-73-48). Princeton: Educational Testing Service.

Murphy, R. T. (1973b). *Investigation of a creativity dimension* (Research Bulletin No. RB-73-12). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1973.tb01027.x

Murphy, R. T. (1977). *Evaluation of the PLATO 4 computer-based education system: Community college component*. Princeton: Educational Testing Service.

Powers, D. E. (1973). *An evaluation of the new approach method* (Program Report No. PR-73-47). Princeton: Educational Testing Service.

Powers, D. E. (1974). *The Virginia Beach extended school year program and its effects on student achievement and attitudes—First year report* (Program Report No. PR-74-25). Princeton: Educational Testing Service.

Powers, D. E. (1975a). *Dual audio television: An evaluation of a six-month public broadcast* (Program Report No. PR-75-21). Princeton: Educational Testing Service.

Powers, D. E. (1975b). *The second year of year-round education in Virginia Beach: A follow-up evaluation* (Program Report No. PR-75-27). Princeton: Educational Testing Service.

Rosenfeld, M. (1973). *An evaluation of the Orchard Road School open space program* (Program Report No. PR-73-14). Princeton: Educational Testing Service.

Shipman, V. C. (1970). *Disadvantaged children and their first school experiences* (Vol. 1, Program Report No. PR-70-20). Princeton: Educational Testing Service.

Shipman, V. C. (1974). *Evaluation of an industry-sponsored child care center. An internal ETS report prepared for Bell Telephone Laboratories. Murray Hill, NJ.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Sigel, I. E. (1976). *Developing representational competence in preschool children: A preschool educational program. In Basic needs, special needs: Implications for kindergarten programs. Selected papers from the New England Kindergarten Conference, Boston*. Cambridge, MA: The Lesley College Graduate School of Education.

Swinton, S., & Amarel, M. (1978). *The PLATO elementary demonstration: Educational outcome evaluation* (Program Report No. PR-78-11). Princeton: Educational Testing Service.

Thomas, I. J. (1970). *A bilingual and bicultural model early childhood education program. Fountain Valley School District title VII bilingual project*. Berkeley: Educational Testing Service.

Thomas, I. J. (1973). *Mathematics aid for disadvantaged students*. Los Angeles: Educational Testing Service.

Trismen, D. A. (1968). *Evaluation of the Education through Vision curriculum—Phase 1*. Princeton: Educational Testing Service.

Trismen, D. A. (with T. A. Barrows). (1970). *Brevard County project: Final report to the Brevard County (Florida) school system* (Program Report No. PR-70-06). Princeton: Educational Testing Service.

Trismen, D. A., Waller, M. I., & Wilder, G. (1976). *A descriptive and analytic study of compensatory reading programs* (Vols. 1 & 2, Program Report No. PR-76-03). Princeton: Educational Testing Service.

Vale, C. A. (1975). *National needs assessment of educational media and materials for the handicapped* [Proposal submitted to Office of Education]. Princeton: Educational Testing Service.

Ward, W. C., & Frederiksen, N. (1977). *A study of the predictive validity of the tests of scientific thinking* (Research Bulletin No. RB-77-06). Princeton: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1977.tb01131.x

Wasdyke, R. G. (1976, August). *An evaluation of the Maryland Career Information System* [Oral report].

Wasdyke, R. G. (1977). *Year 3—Third party annual evaluation report: Career education instructional system project. Newark School District. Newark, Delaware*. Princeton: Educational Testing Service.

Wasdyke, R. G., & Grandy, J. (1976). *Field evaluation of Manhattan Community School District #2 environmental education program*. Princeton: Educational Testing Service.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

Woodford, P. E. (1975). *Pilot project for oral proficiency interview tests of bilingual teachers and tentative determination of language proficiency criteria* [Proposal submitted to Illinois State Department of Education]. Princeton: Educational Testing Service.

Chapter 12
Contributions to Education Policy Research

Richard J. Coley, Margaret E. Goertz, and Gita Z. Wilder

Since Educational Testing Service (ETS) was established in 1947, research has been a prominent gene in the organization's DNA. Nine days after its first meeting, the ETS Board of Trustees issued a statement on the new organization. "In view of the great need for research in all areas and the long-range importance of this work to the future development of sound educational programs, it is the hope of those who have brought the ETS into being that it may make fundamental contributions to the progress of education in the United States" (Nardi 1992, p. 22). Highlighting the important role of research, ETS's first president Henry Chauncey recalled, "We tried out all sorts of names. 'Educational Testing Service' has never been wholly satisfactory because it does leave out the research side" (Nardi 1992, p. 16).

As part of its nonprofit mission, ETS conducts and disseminates research to advance quality and equity in education. Education policy research at ETS was formally established with the founding of the Education Policy Research Institute (EPRI) some 40 years ago, and since then ETS research has focused on promoting equal educational opportunity for all individuals, including minority and educationally disadvantaged students, spanning infancy through adulthood. The major objectives of this work are to provide useful and accurate information on educational opportunity and educational outcomes to the public and to policy makers, to inform the debate on important education issues, and to promote equal educational opportunity for all.

R.J. Coley (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: richardjcoley@gmail.com

M.E. Goertz
University of Pennsylvania, Philadelphia, PA, USA

G.Z. Wilder
Princeton, NJ, USA

R.E. Bennett, M. von Davier (eds.), *Advancing Human Assessment*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-58689-2_12

The purpose of this chapter is to describe ETS's contribution to education policy research. The authors faced three main challenges in accomplishing this goal. First, we had to define what we mean by education policy research. We broadly defined this term to mean work serving to: define the nature of an educational problem that can be addressed by public or institutional policy (e.g., the achievement gap or unequal access to educational opportunities); identify the underlying causes of the problem; or examine the design, implementation, and impact of public or institutional policies or programs designed to address the problem (see, for example, AERA's *Handbook on Education Policy Research* by Sykes et al. 2009).

The second challenge was organizing the work that ETS has conducted. That research has covered three major areas, which were used to select and classify the work described in this chapter. While these areas do not capture the entire scope of ETS's education policy research, they provide important lenses through which to describe that work. The three major areas are:

- Analyzing, evaluating, and informing public policy in educational governance, including school finance; teacher policy; and federal, state, and local education policy.
- Examining differential access to educational opportunity in three areas of long-standing interest to ETS: the gender gap, advanced placement programs, and graduate education.
- Reporting on the educational outcomes of the U.S. population and describing the contexts for these outcomes and for the gaps in outcomes that exist among segments of the population.

The third challenge was selecting from the thousands of research studies that ETS staff have produced over more than half a century. An unfiltered search of ETS ReSEARCHER,[1] a database of publications by ETS staff members, produced nearly 9,000 publications. And while even this database is incomplete, its size is indicative of the scope of the organization's work in psychometrics, statistics, psychology, and education.

Over the past 40 years, the majority of ETS's education policy research was conducted under three organizational structures that operated at different times within the Research and Development division or its predecessors. EPRI was established at ETS in the early 1970s. Its work was expanded in the Education Policy Research division that existed during the 1980s and 1990s. In 1987, the ETS Board of Trustees established the Policy Information Center to inform the national debate on important education policy issues. Hundreds, if not thousands, of projects were conducted and reports produced within these organizational units. The Policy Information Center alone has produced more than 150 policy reports and other publications. These units and their work were heavily supported by internal funds, made possible by the organization's nonprofit status and mission. The organization's financial

[1] The ETS ReSEARCHER database (http://1340.sydneyplus.com/Authors/ETS_Authors/portal.aspx) is available to anyone interested in additional contributions made by the organization to education policy research and to research in measurement, psychology, statistics, and other areas.

commitment to education policy research has been, and continues to be, substantial.

Given this voluminous output, the authors applied the definition of education policy research and the areas described above to assemble what should be considered only a sample. That is, the work described here is reflective of this large body of work, but necessarily incomplete.

Many of ETS's other activities that are education-policy related and contribute to the field of education are not within the scope of this chapter. Some of this important work serves clearinghouse and collaboration functions. An important example includes the networking activities of the Policy Evaluation and Research Center, which collaborates with organizations such as the Children's Defense Fund and the National Urban League and its affiliates to convene a variety of stakeholders around issues related to the achievement gap. These conferences have focused on the particular challenges facing women and girls, the special circumstances of young Black males, issues related to the community college system, and the importance of family factors in students' success in school.

ETS has also had many long-standing relationships with important organizations such as Historically Black Colleges and Universities, the ASPIRA Association, and the Hispanic Association of Colleges and Universities. ETS researchers, in collaboration with the American Association of Community Colleges, examined a number of challenges faced by community colleges in effectively managing both their academic and vocational functions in the context of rapidly changing economic and demographic patterns and the rapid expansion of nondegreed, credentialing, and certification programs (Carnevale and Descrochers 2001). A second example is the Commission on Pathways through Graduate School and into Careers, led by the Council of Graduate Schools and ETS, which resulted in two important reports that identified the major enrollment, retention, and financial issues facing graduate education in the United States (Wendler et al. 2010, 2012).

ETS's policy research has had influence at several levels. It has played important roles in the development of government and institutional policy, in debates about how U.S. students are achieving and the context around student learning, in school and classroom practice, in assessing the status of the nation's human capital, in the shape of internal ETS programs and services, and in the lives of individuals that have been the focus of ETS's work.

In the next section, the first of three major areas, education policy and governance, is reviewed.

12.1 Education Policy and Governance

Over the years, ETS research in this area has covered school finance and governance, teacher policy, and monitoring education policy developments. Each of these areas will be briefly illustrated.

12.1.1 School Finance and Governance

In 1965, University of Chicago sociologist James Coleman led a team that produced the Coleman report, which shed light on unequal schooling conditions and educational opportunities in the United States (Coleman et al. 1966; see also Barone and Beaton, Chap. 8, this volume). At the same time, scholars began to examine how states' funding of elementary and secondary education contributed to these inequities and to raise questions about the constitutionality of these funding systems. ETS researchers played a major role in the subsequent school finance reform movement of the 1970s and 1980s. ETS undertook groundbreaking research on the design and effects of federal, state, and local finance systems—research that laid the foundation for challenges to the constitutionality of state school finance formulas, for the design of alternative funding formulas, and for the development of tools to assist policy makers and the public in its quest to create more equitable funding structures.

Joel Berke, the first director of EPRI, provided the statistical analyses relied upon by both majority and minority justices in the landmark U.S. Supreme Court decision in *Rodriquez vs. San Antonio.* When a closely divided Court ruled that school funding inequities did not violate the Equal Protection Clause of the 14th Amendment of the U.S. Constitution, school finance reformers turned to the education clauses of state constitutions and state courts for relief. Berke and his colleagues worked with attorneys, education groups, and commissions in several states to analyze the allocation of state and local education funds under existing formulas, to assess options for change, and to examine the effects of court-ordered reform systems. For example, a series of reports titled *Money and Education,* issued between 1978 and 1981, examined the implementation of New Jersey's Public School Education Act of 1975, which used a new formula designed to address the wealth-based disparities in education funding declared unconstitutional by the New Jersey Supreme Court (Goertz 1978, 1979, 1981). These reports, along with a follow-up study in the late 1980s, found that although the state increased its education funding, the law fell far short of equalizing expenditures between poor and wealthy communities. These analyses, along with expert testimony by ETS researcher Margaret Goertz, contributed to the New Jersey Supreme Court's 1990 decision in *Abbott v. Burke* to declare the law unconstitutional as applied to the state's poor urban school districts.

ETS staff also worked with policy makers to design new funding formulas in response to court-ordered change. For example, they assisted the New York City Board of Education and the United Federation of Teachers in designing formula adjustments that would address the special financial and educational needs of large urban school systems. The research culminated in *Politicians, Judges, and City Schools* (Berke et al. 1984), a book written to provide New York policy makers with reform options, as well as a better understanding of the political, economic, and social context for reform and of the trade-offs involved in developing a more equitable school finance system.

In addition to policy makers, ETS research has targeted the public. With support from the National Institute of Education and in collaboration with the American Federation of Teachers, ETS researchers sought to demystify the subject of school finance as a way of encouraging informed participation by educators and the general public in school finance debates. While describing school funding formulas in detail, *Plain Talk About School Finance* (Goertz and Moskowitz 1978) also showed that different school finance equalization formulas were mathematically equivalent. Therefore, the authors argued, the selection of a specific formula was secondary to value-laden political decisions about student and taxpayer equity goals for the system, as well as to how to define various components of the formulas (e.g., wealth, taxpayer effort, and student need) and establish the relationships among the components. Building on their analysis of the mathematical properties of school finance formulas, ETS researchers developed the School Finance Equalization Management System (SFEMS), the first generalizable computer software package for financial data analysis and school finance formula simulations (Educational Testing Service 1978a, b). With technical assistance and training from ETS staff, SFEMS was used by nearly a dozen state education agencies and urban school districts to build their capacity to analyze the equity of their state funding systems and to simulate and evaluate the results of different funding approaches.

The wave of legal and legislative struggles over school funding continued throughout the 1980s, and by 1985 more than 35 states had enacted new or revised education aid programs. ETS researchers took stock of this activity in light of the education reform movement that was taking shape in the early 1990s, calling for national standards and school restructuring. *The State of Inequality* (Barton et al. 1991) provided plain talk about school finance litigation and reform, as well as describing how differences in resources available to schools are related to disparities in educational programs and outcomes. The report detailed the disparity in education funding nationally and within states, reviewed data reported by teachers on the connection between instructional resources and student learning, and reviewed a new wave of court rulings on school funding.

School finance research such as that described above focused on disparities in the allocation of resources within states. ETS researchers, however, were among the first to explore disparities *within* school districts, a current focus of school funding debates and policy. In the early 1970s, ETS researcher Joan Baratz examined the implementation of the *Hobson v. Hansen* decision in Washington, DC, which called for the equalization of per-pupil expenditures for all teachers' salaries and benefits within the district. This remedy was designed to address disparities in spending and staffing between schools enrolling many Black and low-income students versus those enrolling many White and affluent students. Baratz (1975) found a significant reduction in the disparity in allocation of all professional staff among the schools as a result of funding equalization. Changes in resources generally involved exchanging highly paid classroom teachers for lower paid teachers, adding teachers in low-spending schools with high pupil/teacher ratios, and redistributing special subject teachers.

A decade later, ETS researchers conducted a congressionally mandated study of school districts' allocation of Title I resources (Goertz 1988). Because most prior research had focused on the distribution of federal funds to local school districts and the selection of schools and students for Title I services, federal policy makers were concerned about the wide range in per-pupil Title I expenditures across school districts and its impact on the delivery of services to students. The ETS study found that variation in program intensity reflected a series of district decisions about how to best meet the needs of students. These decisions concerned program design (e.g., staffing mixes, case loads, settings), type of program (e.g., prekindergarten, kindergarten, bilingual/English as a second language, basic skills replacement), availability and use of state compensatory education funds, and the extent to which allocation decisions reflected differences in student need across Title I schools.

As it is today, the proper organization of responsibility among federal, state, and local governments was a central issue in policy debates in the 1980s about how best to design programs for students with special educational needs. In July, 1981 a team led by ETS researchers began a congressionally mandated study of how federal and state governments interacted as they implemented major federal education programs and civil rights mandates. The study described how states responded to and were affected by federal education programs. Based on analyses of the laws, on case studies conducted in eight states, and interviews with more than 300 individuals at state and local levels, study results portrayed a robust, diverse, and interdependent federal/state governance system. Among the findings was the identification of three broad factors that appeared to explain states' differential treatment of federal programs—federal program signals, state political traditions and climate, and the management and programmatic priorities of state education agencies (Moore et al. 1983).

The topic of school finance was revisited in 2008 when ETS cosponsored a conference, "School Finance and the Achievement Gap: Funding Programs That Work," that explored the relationship between school finance and academic achievement, highlighted programs that successfully close gaps, and examined the costs and benefits of those programs. While much of the discussion was sobering, evidence supporting the cost effectiveness of prekindergarten programs as well as achievement gains made by students in a large urban school district offered evidence that achievement gaps can be narrowed—if the political will, and the money, can be found (Yaffe 2008).

12.1.2 Teacher Policy

While concern about the quality of the nation's teaching force can be traced back to the early twentieth century, during the past 30 years there has been a growing amount of evidence and recognition that teacher quality is a key factor in student achievement. From publication of *A Nation at Risk* in 1983 (National Commission on Excellence in Education [NCEE] 1983), to the National Education Summit in 1989, to the formation of the National Commission on Teaching and America's

Future in 1994, and the No Child Left Behind (NCLB) Act in 2001, teacher quality has remained squarely at the top of national and state education agendas. ETS policy research has responded to the central issues raised about teacher education and teacher quality at various junctures over this period.

12.1.2.1 Research on the Teacher Education Pipeline

Among the areas of education policy that drew significant attention from state policy makers in response to the perceived decline in the quality of the U.S. education system was a focus on improving the preparedness of individuals entering the teaching profession. In the early 1980s, these policies focused on screening program applicants with tests and minimum grade point averages, prescribing training and instruction for those aspiring to become teachers, and controlling access into the profession by requiring aspiring teachers to pass a licensing test or by evaluating a beginning teacher's classroom performance. While the level of state activity in this area was clear, little was known about the substance or impact of these policies. *The Impact of State Policy on Entrance Into the Teaching Profession* (Goertz et al. 1984) identified and described the policies used by states to regulate entrance into the teaching profession and collected information on the impact of these policies.

The study developed and described a *pipeline* model that identified the various points at which state policies can control the entry of individuals into the teaching profession and illustrated the relationships among these points. Next, the study collected information from all 50 states to identify the points of policy intervention and types of policies in effect in each state. In-depth case studies were also conducted in four states to provide details about the political environment and rationale behind the policies, the extent of coordination across policies, and the impact of the policies on teacher supply and equity. While the necessity of screens in the teacher supply pipeline was apparent, the study found that the approaches used by most states were inadequate to address the issues of equity, coordination, and accountability. For example, the study found that screening people out of teaching, rather than developing the talents of those who want to become teachers, is likely to reduce the socioeconomic and racial/ethnic diversity of the nation's teaching force at the very time that schools were becoming more diverse in the composition of their students. The study made recommendations to improve the quality of teachers coming into the profession while recognizing the importance of maintaining a sufficient supply of teachers to staff the nation's increasingly diverse classrooms.

Another movement that took hold during the 1980s in response to criticism directed at traditional teacher education programs was alternate routes to teaching. While these alternate routes took a variety of forms, The Holmes Group (a consortium of education deans from 28 prominent research universities) along with the American Association of Colleges for Teacher Education endorsed the idea of a 5-year teacher education program leading to a master's degree. The idea was that in addition to courses in pedagogy, teachers should have at least the equivalent of an undergraduate degree in the subject they intend to teach. Like the problem, this remedy was not entirely new. In an attempt to understand the likely impact of such an approach, ETS researchers set out to learn about the decades-old master of arts in teaching (MAT) programs, sponsored by the Ford Foundation in response to concerns about the quality of American education generated by the launching of Sputnik. These MAT programs sought to attract bright liberal arts graduates, prepare them for teaching by giving them graduate work in both their discipline and in pedagogy and by providing them with internships in participating school districts.

After searching the Ford Foundation's archives, the researchers put together profiles of the programs and surveyed nearly 1000 MAT program graduates from 1968 to 1969 to see what attracted them to the programs and to teaching, what were their careers paths, and what were their impressions of their preparation. Remarkably, 81% of the MAT program graduates responded to the survey. Among the results: Eighty-three percent entered teaching and one third who entered teaching were still teaching at the time of the survey. Among those who left teaching, the average time teaching was 5 years. Many of the former teachers pursued education careers outside of the classroom. The study, *A Look at the MAT Model of Teacher Education and Its Graduates: Lessons for Today*, concluded that the MAT model was a viable alternative to increase the supply and quality of the nation's teachers, although more modern programs should be designed to recognize the changing composition of the nation's school population (Coley and Thorpe 1985).

A related focus of ETS research during this period was on finding ways to increase the supply of minority teachers. Declining numbers of minority teachers can be attributed to the limited number of minority students entering and completing college, declining interest in education careers, and the policy screens identified in the study described earlier, including the teacher testing movement. *Characteristics of Minority NTE Test-Takers* (Coley and Goertz 1991) sought to inform interventions to increase minority representation in teaching by identifying the characteristics of minority students who met state certification requirements. The study was the first to collect information on candidates' demographic, socioeconomic, and educational background; education experience in college and graduate school; experiences in teacher education programs; career plans and teaching aspirations; and reasons for taking the certification test. The data analyses focused on determining whether successful and unsuccessful National Teachers Examination (NTE) candidates differed significantly on these background and educational characteristics. Four implications drawn were noteworthy. First, many of the minority candidates were the first generation in their families to attend college, and institutions must develop support programs geared to the academic and financial needs of these students. Second, in general, many low socioeconomic status (SES) students who succeeded in college passed the test. Colleges can and do make a difference for disadvantaged students. Third, recruiting and training policies should reflect the large number of minority students who take various routes into and out of teaching. Last, because only half of successful minority candidates planned to make teaching their career, changes to the structure of the teaching profession should be considered, and the professional environment of teaching should be improved to help retain these students.

A recent study by ETS researchers found that minorities remain underrepresented in the teaching profession and pool of prospective teachers (Nettles et al. 2011). The authors analyzed the performance of minority test takers who took ETS's *PRAXIS*® teacher-certification examinations for the first time between 2005 and 2009 and the relationship of performance with test takers' demographic, socioeconomic, and educational backgrounds, including undergraduate major and undergraduate grade point average (UGPA). They also interviewed students and faculty of teacher education programs at several minority-serving colleges and universities to identify challenges to, and initiatives for, preparing students to pass PRAXIS. The report revealed large score gaps between African American and White teacher candidates on selected *PRAXIS I*® and *PRAXIS II*® tests, gaps as large as those commonly observed on the SAT and *GRE*® tests. Selectivity of undergraduate institution, SES, UGPA, and being an education versus a noneducation major were consistently associated with PRAXIS I scores of African American candidates, particularly in mathematics. Recommendations included focusing on strengthening candidates' academic preparation for and achievement in college and providing students with the other skills and knowledge needed to pass PRAXIS.

ETS research has also informed the debate about how to improve teacher education by examining systems of teacher education and certification outside the United States. *Preparing Teachers Around the World* (Wang et al. 2003) compared teacher education in the United States with the systems in high-performing countries, systematically examining the kinds of policies and control mechanisms used to shape the quality of the teaching forces in countries that scored as well or better than the United States in international math and science assessments. The researchers surveyed the teaching policies of Australia, England, Hong Kong, Japan, Korea, the Netherlands, and Singapore. While no one way was identified that the best performing countries used to manage the teacher pipeline, by and large, they were able to control the quality of individuals who enter teacher education programs through more rigorous entry requirements and higher standards than exist in the United States. One of the most striking findings was that students in these countries are more likely to have teachers who have training in the subject matter they teach. And while much has been made in the United States about deregulating teacher education as a way to improve teacher quality, every high-performing country in the study employed significant regulatory controls on teaching, almost all more rigorous than what is found in the United States.

12.1.2.2 Research on the Academic Quality of the Teaching Force

ETS researchers have tracked the quality of the nation's teaching force in several studies. *How Teachers Compare: The Prose, Document, and Quantitative Literacy of America's Teachers* (Bruschi and Coley 1999) took advantage of the occupational data collected in the National Adult Literacy Survey (NALS) to provide a rare look at how the skill levels of teachers compare with other adults and with adults in other occupations. The results of this analysis were quite positive. America's teachers, on average, scored relatively highly on all three literacy scales and performed as well as other college-educated adults. In addition, the study found that teachers were a labor-market bargain, comparing favorably with other professionals in their literacy skills, yet earning less, dispelling some negative stereotypes that were gaining ground at the time.

In related work to determine whether the explosion of reform initiatives to increase teacher quality during the 1990s and early 2000s was accompanied by changes in the academic quality of prospective teachers, ETS research compared two cohorts of teachers (1994–1997 and 2002–2005) on licensure experiences and academic quality. *Teacher Quality in a Changing Policy Landscape: Improvements in the Teacher Pool* (Gitomer 2007) documented improvements in the academic characteristics of prospective teachers during the decade and cited reasons for those improvements. These reasons included greater accountability for teacher education programs, Highly Qualified Teacher provisions under the NCLB Act, increased requirements for entrance into teacher education programs, and higher teacher education program accreditation standards.

12.1.2.3 Research on Teaching and Student Learning

ETS policy research has also focused on trying to better understand the connection between teaching and classroom learning. ETS researchers have used the large-scale survey data available from the National Assessment of Educational Progress (NAEP) to provide insight into classroom practice and student achievement. *How Teaching Matters: Bringing the Classroom Back Into Discussions About Teacher Quality* (Wenglinsky 2000) attempted to identify which teacher classroom practices in eighth-grade mathematics and science were related to students' test scores. The research concluded that teachers should be encouraged to target higher-order thinking skills, conduct hands-on learning activities, and monitor student progress regularly. The report recommended that rich and sustained professional development that is supportive of these practices should be widely available.

ETS researchers conducted a similar analysis of NAEP data to identify teachers' instructional practices that were related to higher science scores and then examined the extent to which minority and disadvantaged students had access to these types of instruction. In addition to providing a rich description of the eighth-grade science classroom and its teachers, *Exploring What Works in Science Instruction: A Look at the Eighth-Grade Science Classroom* (Braun et al. 2009) found that two apparently effective practices—teachers doing science demonstrations and students discussing science in the news—were less likely to be used with minority students and might be useful in raising minority students' level of science achievement.

12.1.2.4 Research on Understanding Teacher Quality

Along with the recognition of the importance of teacher quality to student achievement have come a number of efforts to establish a quantitative basis for teacher evaluation. These efforts are typically referred to as value-added models (VAMs) and use student test scores to compare teachers. To inform the policy debate, ETS published a report on the topic. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models* (Braun 2005) offered advice for policy makers seeking to understand both the potential and the technical limitations that are inherent in such models.

Also related to teacher evaluation, ETS partnered with several organizations as part of the National Comprehensive Center for Teacher Quality (NCCTQ) to produce reports aimed at improving the quality of teaching, especially in high-poverty, low-performing, and hard-to-staff schools. One effort by ETS researchers lays out an organizational framework for using evaluation results to target professional development opportunities for teachers, based on the belief that teacher accountability data can also be used to help teachers improve their practice (Goe et al. 2012). To help states and school districts construct high-quality teacher evaluation systems for employment and advancement, Goe and colleagues collaborated with NCCTQ partners to produce a practical guide for education policy makers on key areas to be addressed in developing and implementing new systems of teacher evaluation (Goe et al. 2011).

Work on teacher quality continues as ETS researchers grapple with policy makers' desire to hold teachers accountable for how much students learn. Studies that examine a range of potential measures of teaching quality, including classroom observation protocols, new measures of content knowledge for teaching, and measures based on student achievement, are ongoing. The studies investigate a wide range of approaches to measuring teaching quality, especially about which aspects of teaching and the context of teaching contribute to student learning and success.

12.1.3 Monitoring Education Policy Developments

Much of the Policy Information Center's work has focused on reporting on education policy developments and on analyzing the educational achievement and attainment of the U.S. population, as well as identifying and describing a range of factors that influence those outcomes. In monitoring and describing the changing education policy landscapes that evolved over the decades, the Center sought to anchor data on achievement and attainment to relevant educational reform movements. A sample of that work is provided next.

The decade of the 1980s that began with the publication of *A Nation at Risk* (NCEE 1983) witnessed extensive policy changes and initiatives led by governors and state legislatures, often with strong backing from business. *The Education Reform Decade* (Barton and Coley 1990) tracked changes at the state level between

1980 and 1990 in high school graduation requirements, student testing programs, and accountability systems, as well as sweeping changes in standards for teachers. Changes at the local level included stricter academic and conduct standards, more homework and longer school days, and higher pay for teachers. By the decade's end, 42 states had raised high school graduation requirements, 47 states had established statewide testing programs, and 39 states required passing a test to enter teacher education or begin teaching (Coley and Goertz 1990).

Against this backdrop often referred to as the *excellence movement*, the report provided a variety of data that could be used to judge whether progress was made. These data included changes in student achievement levels, several indicators of student effort, and success in retaining students in school. Data were also provided regarding progress toward increasing equality and decreasing gaps between minority and majority populations and between males and females. Some progress in closing the gaps in achievement, particularly between White and Black students, as well as modest progress in other areas, prompted this November 15, 1990, headline in *USA Today*: "Reforms Put Education on Right Track" (Kelly 1990). Then-ETS President Gregory R. Anrig noted at the press conference releasing the report, "The hallmark of the decade was a move toward greater equality rather than a move toward greater excellence" (Henry 1990, p. 1).

One of the more tangible outcomes of the education-reform decade was the near universal consensus that the high school curriculum should be strengthened. The National Commission on Excellence in Education recommended that all high school students should complete a core curriculum of 4 years of English; 3 years each of social studies, science, and mathematics; 2 years of a foreign language; and one-half year of computer science. Progress toward attaining this new standard was tracked by two ETS reports. *What Americans Study* (Goertz 1989) and *What Americans Study Revisited* (Coley 1994) reported steady progress in student course-taking between 1982 and 1990. While only 2% of high school students completed the core curriculum in 1982, the percentage rose to 19 in 1990. In addition, 40% of 1990 high school graduates completed the English, social studies, science, and mathematics requirements, up from 13% in 1982. The 1994 report also found that the level of mathematics course-taking increased in advanced sequences and decreased in remedial ones.

Along with changes in what students study, the explosion of state testing programs that occurred in the 1970s carried over and expanded in the 1980s with the excellence movement. Perhaps the most notable change was the growth of elementary and secondary school testing across the states. As the 1990s began, there were increasing calls to broaden educational assessment to include performance assessment, portfolios of students' work, and constructed-response for which students had to come up with an answer rather than fill in a bubble. By the 1992–1993 school year, only Iowa, Nebraska, New Hampshire, and Wyoming did not have a state testing program.

Testing in America's Schools (Barton and Coley 1994) documented the testing and assessment changes that were occurring across the country. The report used information from NAEP, a study from what was then the U.S. General Accounting

Office, and a survey of state testing directors conducted by the Council of Chief State School Officers to provide a profile of state testing programs in the early 1990s, as well as a view of classroom testing. The report noted that while the multiple-choice exam was still America's test of choice, the use of alternative methods was slowly growing, with many states using open-ended questions, individual performance assessments, and portfolios or learning records.

As the 1990s drew to a close, President Clinton and Vice President Al Gore called for connecting all of America's schools to the *information superhighway*, federal legislation was directing millions of dollars to school technology planning, and a National Education Summit of governors and business leaders pledged to help schools integrate technology into their teaching. Amid this activity and interest *Computers and Classrooms: The Status of Technology In U.S. Schools* (Coley et al. 1997) was published to meet a need for information on how technology is allocated among different groups of students, how computers are being used in schools, how teachers are being trained in its use, and what research shows about the effectiveness of technology. The report made headlines in *The Washington Post, USA Today, The Philadelphia Inquirer,* and *Education Week* for uncovering differences in computer use by race and gender. Among other findings were that poor and minority students had less access than other students to computers, multimedia technology, and the Internet.

While publications such as *Education Week* now take the lead in describing the policy landscape, there are occasions when ETS research fills a particular niche. Most recently, for example, information on pre-K assessment policies was collected and analyzed in *State Pre-K Assessment Policies: Issues and Status* (Ackerman and Coley 2012). In addition to information on each state's assessments, the report focused on reminding policy makers about the special issues that are involved in assessing young children and on sound assessment practices that respond to these challenges. In this area, ETS contributes by keeping track of important developments while at the same time providing leadership in disseminating tenets of proper test use.

12.2 Access to Educational Opportunities Along the Education Pipeline

ETS's mission has included broadening access to educational opportunities by groups other than the White middle-class population that had traditionally—and often disproportionately—enjoyed the benefits of those opportunities. Increasing access to graduate education, particularly for underrepresented groups, requires improving educational opportunities from early childhood through high school and college. Over the years, ETS researchers have studied differential access to quality education at all points along the educational pipeline. For example, ETS research on early childhood education has included seminal evaluations of the impact on traditionally underserved groups of such educational television programs as Sesame Street and The Electric Company (Ball and Bogatz 1970; Ball et al. 1974), and improving the quality of early childhood assessments (Ackerman and Coley 2012; Jones 2003). Other researchers have focused on minority students' access to mathematics and science in middle schools (see, for example, Clewell et al. 1992), and individual and school factors related to success in high school (see, for example, Ekstrom et al. 1988). ETS research on the access of underrepresented groups to higher education has also included evaluations of promising interventions, such as the Goldman Sachs Foundation's Developing High Potential Youth Program (Millett and Nettles 2009). These and other studies are too numerous to summarize in this chapter. Rather, we focus on contributions of ETS research in several areas of longstanding interest to the organization—gender differences, access to advanced placement courses in high school, and access to graduate education.

12.2.1 The Gender Gap

Much has been written about the gender gap. ETS has traditionally tracked the trajectories of scores on its own tests, and multiple reports have been dedicated to the topic. A 1989 issue of *ETS Policy Notes* examined male-female differences in NAEP results and in SAT and *PSAT/NMSQT*® scores (Coley 1989). An entire volume by Warren W. Willingham and Nancy Cole was devoted to the topic in the context of test fairness (Willingham and Cole 1997). And a 2001 report deconstructed male-female differences within racial/ethnic groups along with course-taking data, attempting to understand gender differences in educational achievement and opportunity across racial/ethnic groups (Coley 2001). The consensus from much of this work has been that the causes of the male-female achievement gap are many, varied, and complex.

In 1997, then-president of ETS Cole authored a report titled *The ETS Gender Study: How Males and Females Perform in Educational Settings* (Cole 1997). The report was based on 4 years of work by multiple researchers using data from more than 1500 data sets, many of them large and nationally representative. The collective studies used 400 different measures that cut across grades, academic subjects, and years and involved literally millions of students.

Although the study yielded many important and interesting findings, Cole chose to focus on several that were contrary to common expectations. Among them were the following:

- For many subjects, the differences between males and females are quite small, but there are some real differences in some subjects.
- The differences occur in both directions. In some areas, females outperform males, and in others the opposite is true.

- Dividing subjects by component skills produces a different picture of gender differences than those found for academic disciplines more generally.
- Gender differences increase over years in school. Among fourth-grade students, there are only minor differences in test performance on a range of school subjects. The differences grow as students progress in school and at different rates for different subjects.
- Gender differences are not easily explained by single variables such as course-taking or types of test. They are also reflected in differences in interests and out-of-school activities.

Cole concluded that "…while we can learn significant things from studying group behavior, these data remind us to look at each student as a unique individual and not stereotype anyone because of gender or other characteristics" (Cole 1997, p. 26).

Over the years, ETS researchers have sought to determine what factors contribute to the underrepresentation of women in the fields of science, technology, engineering, and mathematics (STEM), going back to elementary and secondary education. Marlaine E. Lockheed, for example, conducted studies of sex equity in classroom interactions (Lockheed 1984) and early research on girls' participation in mathematics and science and access to technology (Lockheed 1985; Lockheed et al. 1985). Building on this and related work, Clewell et al. (1992) identified what they determined were major barriers to participation by women and minorities in science and engineering: (a) negative attitudes toward mathematics and science; (b) lower performance levels than White males in mathematics and science courses, and on standardized tests; (c) limited exposure to extracurricular math- and science-related activities, along with failure to a participate in advanced math and science courses in high school; and (d) lack of information about or interest in math or science careers. Making a case for developing interventions aimed at the critical middle school years, they offered descriptions and case studies of ten intervention programs, then relatively recent phenomena, that the authors considered successful, along with a series of recommendations derived from the programs.

12.2.2 Access to Advanced Placement®

Providing high school students access to advanced coursework has long been considered an important means of preparing students for future success. This preparation is particularly important for minority students, who score, on average, lower than nonminority students. ETS researchers studied the characteristics of minority students with high SAT scores and found that these students tended to excel in advanced coursework in high school, including advanced placement courses (Bridgeman and Wendler 2005).

The College Board's *Advanced Placement Program*® (*AP*®) is a collaborative effort between secondary and postsecondary institutions that provides students