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 policy-

makers 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 profi-

ciency 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.