

## Base-Year to First Follow-up Cognitive Test Battery<sup>1</sup>

### 2.3.1 Base-Year Reading and Mathematics Assessments

The purpose of the ELS:2002 assessment battery is to provide measures of student achievement in reading and mathematics that can be related to student background variables and educational processes, for individuals and for population subgroups. The reading and mathematics tests must provide accurate measurement of the status of individuals at a given point in time. The mathematics test must provide accurate measurement of their cognitive growth over time. Assessment data in ELS:2002 will be used to study factors that contribute to individual and subgroup differences in achievement.

#### 2.3.1.1 Test Design and Format

Test specifications for ELS:2002 were adapted from frameworks used for NELS:88. Math tests contained items in arithmetic, algebra, geometry, data/probability, and advanced topics and were divided into process categories of skill/knowledge, understanding/comprehension, and problem solving. Through inclusion of PISA items, the ELS:2002 math tests placed a somewhat greater emphasis on practical applications and problem solving than did the NELS:88 test forms. Reading tests consisted of reading passages of one paragraph to one page in length, followed by three to six questions based on each passage. The reading passages included literary material as well as topics in the natural and social sciences. Several passages require interpretation of graphs. Questions were categorized as reproduction of detail, comprehension, or inference/evaluation. The test questions were selected from previous assessments: NELS:88, NAEP, and PISA. Most, but not all, were multiple choice. The number of items in each stage of the test is indicated in table 4.

**Table 4. Number of items in each ELS:2002 base-year test form for assessing achievement in mathematics and reading, by test form: 2002**

Form	Mathematics	Reading
Routing test	15	14
Second stage tests		
Form X (low difficulty)	25	16
Form Y (middle difficulty)	27	17
Form Z (high difficulty)	27	15
Form V (single stage: limited time, mathematics only)	23	†

† Not applicable.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002), "Base Year, 2002."

<sup>1</sup> Taken from Ingels, S.J., Pratt, D.J., Rogers, J.E., Siegel, P.H., and Stutts, E.S. (2005). *Education Longitudinal Study of 2002: Base-Year to First Follow-up Data File Documentation* (NCES 2006-344). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

### 2.3.2 First Follow-up Assessment

This section describes the development and format of the ELS:2002 first follow-up mathematics assessment, scoring procedures, score descriptions, and summary statistics. It includes a discussion of links (through equating or concordance) with other studies (NELS:88 and PISA:2000).

The purpose of the ELS:2002 assessments is to provide measures of student achievement in reading and mathematics that can be related to student background variables and educational processes, for individuals and for population subgroups. Reading<sup>2</sup> and mathematics assessments were conducted in the sophomore base year; in the first follow-up, only a mathematics test was administered. Assessment data in ELS:2002 will be used to study factors associated with individual and subgroup differences in achievement. The reading and mathematics tests must provide accurate measurement of the status of individuals at a given point in time and, for mathematics, must provide accurate measurement of their cognitive growth over time.

#### 2.3.2.1 Test Design and Format

As with the base-year test design, the specifications for the ELS:2002 first follow-up math test were adapted from frameworks used for NELS:88. Mathematics tests contained items in arithmetic, algebra, geometry, data/probability, and advanced topics and were divided into process categories of skill/knowledge, understanding/comprehension, and problem solving. However, like the base-year test, the ELS:2002 mathematics tests placed a greater emphasis on practical applications and problem solving than did the NELS:88 tests. The test questions were selected from previous assessments: NELS:88, NAEP, and PISA. Items were field tested<sup>3</sup> one year prior to the 10th- and 12th-grade surveys, and some items were modified based on field test results. Final forms were assembled based on psychometric characteristics and coverage of framework categories. In the base year, about 10 percent of math questions were open ended; all of the 12th-grade mathematics questions were presented in multiple-choice format.

The ELS:2002 assessments were designed to maximize the accuracy of measurement that could be achieved in a limited amount of testing time, while minimizing floor and ceiling effects, by matching sets of test questions to initial estimates of students' achievement. (For definitions of floor effects, ceiling effects, and other technical terms, see the glossary in appendix E.) In the base year, this was accomplished by means of a two-stage test. In 10th grade, all students received a short multiple-choice routing test, scored immediately by survey administrators who then assigned each student to a low, middle, or high difficulty second-stage form, depending on the student's number of correct answers in the routing test. In the 12th-grade administration, students were assigned to an appropriate test form based on their performance in 10th grade. Cut points for the 12th-grade low, middle, and high forms were calculated by pooling

---

<sup>2</sup> Please refer to base-year documentation (Ingels et al. 2004, NCES 2004–405) for additional information on the 10th-grade reading test.

<sup>3</sup> For more details about the field tests, see Burns et al. (2003) (NCES 2003–03) and appendix J of this manual.

**Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25**

information from the field tests for 10th and 12th grades in 2001, the 12th-grade field test in 2003, and the 10th-grade national sample. Item and ability parameters were estimated on a common scale. Growth trajectories for longitudinal participants in the 2001 and 2003 field tests were calculated, and the resulting regression parameters were applied to the 10th-grade national sample. Test forms were designed to match the projected achievement levels of the lowest and highest 25 percent, and the middle 50 percent, of the base-year sample 2 years later. An additional test form with a broad range of item difficulty was assembled for administration to follow-up participants who were new to the sample or who had not received a math score in 10th grade. Additions to and deletions from the base-year sample resulted in 23 percent, 42 percent, and 26 percent of the follow-up sample taking the low, middle, and high difficulty forms, respectively, with the remaining 10 percent taking the broad-band form. Each of the four test forms contained 32 multiple-choice items.

**2.3.2.2 Scoring Procedures**

Eleven test records were deleted because tests were incomplete (fewer than 10 items answered) and six more because regular response patterns (e.g., all answers were “A,” or “ABCABCABC...,” etc.) indicated lack of motivation to answer questions to the best of the student’s ability.

The scores used to describe students’ performance on the direct cognitive assessment are broad-based measures that report performance as a whole. The scores are based on IRT, which uses patterns of correct, incorrect, and omitted answers to obtain ability estimates that are comparable across different test forms.<sup>4</sup> In estimating a student’s ability, IRT also accounts for each test question’s difficulty, discriminating ability, and a guessing factor.

IRT has several advantages over raw number-right scoring. By using the overall pattern of right and wrong responses to estimate ability, IRT can compensate for the possibility of a low-ability student guessing several difficult items correctly. If answers on several easy items are wrong, a correct difficult item is assumed, in effect, to have been guessed. Omitted items are also less likely to cause distortion of scores, as long as enough items have been answered right and wrong to establish a consistent pattern. Unlike raw number-right scoring, which necessarily treats omitted items as if they had been answered incorrectly, IRT procedures use the pattern of responses to estimate the probability of correct responses for all test questions. Finally, IRT scoring makes it possible to compare scores obtained from test forms of different difficulty. The common items present in overlapping forms and in overlapping administrations (10th grade and 12th grade) allow test scores to be placed on the same scale.

In the first follow-up survey, IRT procedures were used to estimate longitudinal gains in achievement over time by using common items present in both the 10th- and 12th-grade forms. Items were pooled from both the 10th- and 12th-grade administrations and anchored to the IRT scale of the NELS:88 survey of 1988–92. Item parameters were fixed at NELS:88 values for the items that had been taken from the NELS:88 test

---

<sup>4</sup> For an account of Item Response Theory, see Hambleton (1989) or Hambleton, Swaminathan, and Rogers (1991).

## Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25

battery and to base-year values for non- NELS:88 items. In each case, the fit of the follow-up item response data to the fixed parameters was evaluated, and parameters for common items whose current performance did not fit previous patterns were reestimated, along with non-NELS:88 items new to the follow-up tests.

### 2.3.2.3 Score Descriptions and Summary Statistics

Several different types of scores that can be used to describe students' performance on the cognitive assessment are described in detail below. IRT-estimated number-right scores measure students' performance on the whole item pool. NELS:88-equated number-right scores estimate how a student would have performed on the 1992 mathematics scale of NELS:88. Standardized scores (T-scores) report students' performance relative to their peers. Quartile scores divide the estimated population distributions for convenience in analyzing relationships of cognitive skills with other variables. NELS:88-equated proficiency probabilities estimate the probability that a given student would have demonstrated proficiency for each of the five mathematics levels defined for the NELS:88 survey in 1992.<sup>5</sup>

**IRT-estimated number right.** These scores are estimates of the number of items students would have answered correctly if they had responded to all of the 85 questions in the mathematics item pool (i.e., all items that appeared on any of the mathematics forms) in the 10th- and 12th-grade administrations combined. The ability estimates and item parameters derived from the IRT calibration can be used to calculate each student's probability of a correct answer for each of the items in the pool. These probabilities are summed to produce the IRT- estimated number-right scores. These scores are not integers because they are sums of probabilities, not counts of right and wrong answers.

It is important to note that the item pool for base-year and first follow-up mathematics forms combined differs from the sophomore-only pool used to report scale scores in the base year. The combined sophomore forms contained a total of 72 items, with 13 additional, harder items added in 12th grade to extend the range of the scale. To place base-year and first follow-up scores on the same scale so that gains over time can be measured, the base-year IRT- estimated number-right scores have been replaced with scores on the new 85-item combined scale. Table 5 shows variable names, descriptions, and summary statistics for the IRT-estimated number-right scores in the new metric that applies to both rounds of the survey. First follow-up statistics are reported both for all first follow-up participants and for the subset of students who were in 12th grade at the time of the survey. The samples include all students with test scores, as well as imputed scores for students who were not tested. (For a discussion of imputation in ELS:2002, see chapter 3 of this manual). The reliability of the test scores

---

<sup>5</sup> For further information on the NELS:88 proficiency levels, see Rock and Pollack (1995a), *Psychometric Report for the NELS:88 Base Year Through Second Follow-up* (NCES 95-382). For examples of the use of the NELS:88- equated probability proficiency scores in the context of cross-sectional estimation of status in ELS:2002, see chapter 5 of Ingels et al. (2005), *A Profile of the American High School Sophomore in 2002* (NCES 2005-338). For examples of longitudinal use of the probability of proficiency scores (in NELS:88), see chapter 4 of Scott et al. (1995), *Two Years Later: Cognitive Gains and School Transitions of NELS:88 Eighth Graders* (NCES 95-436).

Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25

is a function of the variance of repeated estimates of the IRT ability parameter (within-variance), compared with the variability of the whole sample. For the combined base-year and first follow-up mathematics tests, the reliability was 0.92. This applies to all scores derived from the IRT estimation, including the standardized and quartile scores.

**Table 5. Item Response Theory (IRT)-estimated number-right mathematics scores in 85-item metric: 2004**

Variable name	Description	Range	Weighted mean	Weighted standard deviation
BYTXMIR2	Mathematics IRT-estimated number right, base year, reestimated on longitudinal scale	0-85	42.2	14.0
F1TXMIR2	Mathematics IRT-estimated number right, longitudinal scale, all first follow-up participants	0-85	48.3	15.1
F1TXMIR2	Mathematics IRT-estimated number right, longitudinal scale, first follow-up participants who were in 12th grade	0-85	48.6	15.1

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002), “Base Year, 2002” and “First Follow-up, 2004.”

**Standardized scores (T-scores).** T-scores provide norm-referenced measurements of achievement; that is, estimates of achievement level relative to the population as a whole. A high mean T-score for a particular subgroup indicates that the group’s performance is high in comparison with other groups. It does not represent mastery of a particular set of skills, only that the subgroup’s mastery level is greater than a comparison group. In other words, T-scores provide information on status compared to students’ peers, whereas the IRT-estimated number- right scores represent status with respect to achievement on a particular criterion set of test items. The T-scores can only provide an indicator of the extent to which an individual or a subgroup ranks higher or lower than the national average. The standardized score reported in the database (F1TXMSTD) is a transformation of the IRT theta (ability) estimate, rescaled to a mean of 50 and standard deviation of 10.

**Quartile scores.** Quartile scores (BYTXMQU) divide the sample into four equal groups, based on the weighted distribution of mathematics scores. Quartile 1 corresponds to the lowest achieving quarter of the population, quartile 4 to the highest.

**2.3.2.4 Links to Other Surveys**

Scores for ELS:2002 first follow-up are reported on scales that permit comparisons with mathematics data for NELS:88 12th-graders in 1992. In addition, ELS:2002 base-year mathematics scores were linked to the 2003 PISA mathematics scale. (In the base year, ELS:2002 reading scores were put on the PISA [2000] literacy scale; for details see Ingels et al. [2004], NCES 2004–405.) The link to the NELS:88 scales represents a “true” equating. This means that the tests may be considered interchangeable or, in other words, a score on one exam should be equivalent to a score on the other exam. Several conditions must be met for equating two tests. Most important, the tests must

Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25

measure the same content. Similarity of format, length, reliability, and subgroup performance also supports the interpretation of interchangeable scores.

**PISA concordance.** The ELS:2002 and PISA mathematics tests did not share enough items to permit common-item equating, so score scales were linked by means of *equipercentile equating* (see Ingels et al. [2004] for a description of the reading equating with PISA). If two exams measuring the same construct are given to two samples from the same population, the score corresponding to a certain percentile on one exam may be considered to be equivalent to the score on the other exam that represents the same percentile of the population. ELS:2002 and PISA test instruments, scoring methods, and populations differed in several respects that affect the equating procedures and interpretation of linked scores.

The most important difference between PISA and ELS:2002 is the definition of the population sampled in each case. Equipercentile equating assumes that the two samples being equated come from the same population. However, important differences exist between PISA and ELS:2002 (see table 6). The PISA population was based on age (students born in 1987), whereas ELS:2002’s population was based on grade (high school sophomores). Although the spring term administration dates for PISA and ELS:2002 overlapped, the range of PISA dates was later in the school year, suggesting the possibility of higher scores due to additional weeks or months of schooling.

**Table 6. ELS:2002 and Program for International Student Assessment: Spring 2003 (PISA:spring 2003), by sample characteristics: 2002 and 2003**

ELS:2002 sample	PISA:spring 2003 sample
10th-graders only	Different grades
Different ages; modal age = 15	Ages 15.25 to 16.25 years
Testing began in January 2002	Testing began in April 2003
14,543 tested	3,983 tested

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002); and Program for International Student Assessment (PISA).

Because of these differences, subsamples of each group were used to compute equivalent percentiles (see table 7). Transformations were computed based on the 10th-graders from each survey who were within a specified range of ages and testing dates. The resulting transformation was then applied to all ELS:2002 students. To make the PISA sample more nearly equivalent to the ELS:2002 sample, only PISA 10th-graders were used in the equating subsample. To make the ELS:2002 sample more nearly equivalent to the PISA sample, only ELS:2002 students between the ages of 15.25 years and 16.25 years (the approximate age range for PISA examinees) were used. ELS:2002 students who were tested before March 15 or after May 31 were deleted from the equating sample. The restricted samples were intended to be representative of 10th-graders between the ages of 15.25 and 16.25 years.

Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25

**Table 7. ELS:2002 and Program for International Student Assessment: Spring 2003 (PISA:spring 2003) equating sample: 2002 and 2003**

ELS:2002 equating sample	PISA:spring 2003 equating sample
10th-graders only	10th-graders only
15.25- to 16.25-year-olds	15.25- to 16.25-year-olds
Exams offered from March 15 to May 31	Exams offered from April 1 to May 31
Equating sample N = 2,743	Equating sample N = 2,400

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002); and Program for International Student Assessment (PISA).

The equipercentile equating was carried out using five-moment smoothing of the weighted frequency distributions. Plots of the equipercentile-equated scores showed extreme deviations in the tails of the distribution from a trend line based on linear approximation. These deviations are probably due to the methodology employed in PISA scoring: the PISA scores are transformations of normally distributed IRT ability estimates, which, if no shrinkage is imposed, tend to have long tails. The ELS:2002 scores, which are sums of probabilities, do not. As a result, the equipercentile conversion becomes distorted in the tails of the distributions. Throughout most of the score range, a quarter point difference in ELS:2002 math scale corresponds to a difference of 2 to 3 points in the PISA metric. But, in the extreme tails of the distribution, a quarter point difference in ELS:2002 math score corresponds to a difference of 5 to 10 points or more in the PISA metric. For this reason, the equipercentile equating was carried out without the data in the top and bottom tails of each distribution. Then the equipercentile transformation was used to link the scores for the middle 90 percent of the students, and the remaining scores were linked based on the linear approximation of the equating transformation. The cutoff points for using equipercentile versus linear transformation were selected such that the ELS:2002 to PISA link would be monotonic. Table 8 shows the linking methods for implementing PISA:spring 2003 math scales in ELS:2002.

**Table 8. Linking methods for implementing Program for International Student Assessment: Spring 2003 (PISA:spring 2003) math scales in ELS:2002: 2002 and 2003**

ELS:2002 scale score range	Equating method	Weighted percent of data
12.60–22.05	Linear approximation	10.0
22.06–51.81	Equipercentile transformation	76.0
51.82–68.90	Linear approximation	14.0

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002); and Program for International Student Assessment (PISA).

Data users should keep in mind that the differences between the ELS:2002 and PISA:spring 2003 tests, scoring methods, and populations mean that the link reported here cannot be considered to be a true equating. Although procedures were carried out to compensate for population differences and scoring methods, no claim is made that the scores may be treated as equivalent. It is more appropriate to refer to this link as a concordance: the PISA-scale score represents the score level achieved by students of

**Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25**

the same percentile rank in two populations that were matched as closely as possible given the differences described above.

**NELS:88-equated scores.** Equating the ELS:2002 scale scores to the NELS:88 scale scores was completed through common-item or *anchor equating*. The ELS:2002 and NELS:88 mathematics tests shared 44 math items. These common items provided the link that made it possible to obtain ELS:2002 student ability estimates on the NELS:88 ability scale. (The ELS:2002 data for 12 additional math items did not fit the NELS:88 IRT parameters, so these items were not treated as common items for the purpose of equating.) Parameters for the common items were fixed at their NELS:88 values, resulting in ability estimates consistent with the NELS:88 metric.

The NELS:88-equated IRT-estimated number-right scores for mathematics are estimates of the number of items students would have answered correctly had they taken the NELS:88 exam and responded to all items in the mathematics items pool. The NELS:88 item pool contained 81 mathematics items in all test forms administered in grades 8, 10, and 12. Table 9 shows mathematics scores for ELS:2002 students, reported on the NELS:88 score scale.

**Proficiency probability scores.** The criterion-referenced NELS:88-equated proficiency probability scores are based on clusters of items that mark different levels on the mathematics scale. Clusters of four items were identified in the NELS:88 tests that marked five hierarchical levels in mathematics:

1. simple arithmetical operations on whole numbers;
2. simple operations with decimals, fractions, powers, and roots;
3. simple problem solving requiring the understanding of low-level mathematical concepts;
4. understanding of intermediate-level mathematical concepts and/or multistep solutions to word problems; and
5. complex multistep word problems and/or advanced mathematical material.

The proficiency levels are hierarchical in the sense that mastery of a higher level typically implies proficiency at lower levels. In NELS:88, students were judged to be proficient if three of the four items in a cluster were answered correctly. The NELS:88-equated proficiency probabilities were computed using IRT-estimated item parameters calibrated in NELS:88. Each proficiency probability represents the probability that a student would pass a given proficiency level defined as above in the NELS:88 sample.

Table 9 shows variable names, descriptions, and summary statistics for the NELS:88-equated proficiency probability scores.

**Table 9. ELS:2002 Item Response Theory (IRT) National Education Longitudinal Study of 1988 (NELS:88)-equated estimated number-right score and proficiency probability scores: 2004**

Variable name	Description	Range	Weighted mean	Weighted standard deviation
F1NELS2M	thematics—NELS-equated estimated number right (1992 scale)	0–81	50.10	14.20
F1TX1MPP	Mathematics—level 1	0–1	.96	.12
F1TX2MPP	Mathematics—level 2	0–1	.78	.37
F1TX3MPP	Mathematics—level 3	0–1	.62	.45
F1TX4MPP	Mathematics—level 4	0–1	.35	.41
F1TX5MPP	Mathematics—level 5	0–1	.04	.14

SOURCE: U.S. Department of Education, National Center for Education Statistics, Education Longitudinal Study of 2002 (ELS:2002), “First Follow-up, 2004.”

**Choosing the appropriate score for analysis.** The IRT-estimated number-right, standardized scores (T-scores), proficiency, and quartile scores are all derived from the IRT model and are based on all of the student’s responses to the mathematics assessment. That is, the pattern of right and wrong answers, as well as the characteristics of the assessment items themselves, are used to estimate a point on an ability continuum, and this ability estimate, theta, then provides the basis for criterion-referenced and norm-referenced scores. The choice of the most appropriate score for analysis purposes should be driven by the context in which it is to be used.

The *IRT-estimated number-right scores* are overall, criterion-referenced measures of status at a point in time. The criterion is the set of skills defined by the framework and represented by the assessment item pool. These scores are useful in identifying cross-sectional differences among subgroups in overall achievement level. They provide a summary measure of achievement useful for correlational analysis with status variables, such as demographics, school type, or behavioral measures, and may be used in multivariate models as well. These scores may also be used as longitudinal measures of overall growth. However, gains made at different points on the score scale have qualitatively different interpretations. For example, students who made 5- point gains by mastering arithmetical operations are learning very different lessons from those gaining 5 points at the high end of the scale by learning more advanced mathematics. Although the gains in number of scale score points may be the same, the interpretation, and the relationship with other factors such as coursework, can be expected to be quite different. Comparison of gain in scale score points is most meaningful for groups that started with similar initial status.

The *standardized scores (T-scores)* are also overall measures of status at a point in time, but they are norm-referenced rather than criterion-referenced. They do not answer the question, “What skills do students have?” but rather, “How do they compare with their peers?” The transformation to a familiar metric with a mean of 50 and standard deviation of 10 facilitates comparisons in standard deviation units. For example, an individual with a T-score of 65 (or a subgroup with a mean of 65) has demonstrated achievement one and one-half standard deviations above the national average for 12th-graders, whereas a score of 45 would correspond to half a standard deviation below the

**Data Collected Through the ELS:2002 > The ELS:2002 Assessment Battery > Slide 8 of 25**

norm. These numbers do not indicate whether students have mastered a particular body of material, but rather what their standing is relative to others.

*Quartile scores* are convenient normative scores for the user who wishes to focus on analysis of background or process variables separately for students at different achievement levels. For example, one might want to compare the school experiences or educational aspirations of students in the lowest mathematics quartile with those of students in the highest quartile group.

*NELS:88-equated estimated number-right* and *proficiency probability scores* may be used in a number of ways. Because they are calibrated on the NELS:88 scale, they may be used for cross-sectional comparisons of students' mathematics achievement in 2004 compared with their counterparts in 1992. The NELS:88-equated number-right scores reflect performance on the whole pool of 81 NELS:88 mathematics items, whereas the proficiency probability scores are criterion-referenced scores that target a specific set of skills. The mean of a proficiency probability score aggregated over a subgroup of students is analogous to an estimate of the percentage of students in the subgroup who have displayed mastery of the particular skill. The proficiency probability scores are particularly useful as measures of gain, because they can be used to relate specific treatments (such as selected coursework) to changes that occur at different points along the score scale. For example, two groups may have similar gains in total scale score points, but for one group, gain may take place at an upper skill level, and for another, at a lower skill level. One would expect to see a relationship between gains in probability of proficiency at a particular level and curriculum exposure, such as taking mathematics courses relevant to the skills being mastered.

## References

- Burns, L.J., Heuer, R., Ingels, S.J., Pollack, J.M., Pratt, D.J., Rock, D., Rogers, J., Scott, L.A., Siegel, P., and Stutts, E. (2003). *ELS:2002 Base Year Field Test Report* (NCES Working Paper 2003–03). U.S. Department of Education. Washington, DC: National Center for Education Statistics.
- Hambleton, R.K. (1989). Principles and Selected Applications of Item Response Theory. In R.L. Linn (Ed.), *Educational Measurement*, 3rd ed. (pp. 147–200). New York: MacMillan.
- Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
- Ingels, S.J., Burns, L.J., Chen, X., Cataldi, E.F., and Charleston, S. (2005). *A Profile of the American High School Sophomore in 2002* (NCES 2005–338). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.
- Ingels, S.J., Pratt, D.J., Rogers, J., Siegel, P.H., and Stutts, E.S. (2004). *Education Longitudinal Study of 2002: Base Year Data File User's Manual* (NCES 2004–405). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available: <http://nces.ed.gov/pubsearch>.
- Rock, D.A., and Pollack, J.M. (1995a). *Psychometric Report for the NELS:88 Base Year Through Second Follow-up* (NCES 95–382). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.
- Scott, L.A., Rock, D.A., Pollack, J.M., and Ingels, S.J. (1995). *Two Years Later: Cognitive Gains and School Transitions for NELS:88 Eighth Graders* (NCES 95–436). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.