2. Measurement and Analysis Considerations

Although appendix A provides details on many aspects of the NLTS2 design, measurement, and analysis approaches, this chapter highlights the following, which are particularly important in helping readers interpret the findings reported in subsequent chapters:

- the research versions of the direct assessment subtests;
- determining the type of assessment to be administered;
- assessment procedures;
- analysis approaches; and
- the population of youth with disabilities to whom the findings generalize.

Research Versions of the Woodcock-Johnson III Subtests

As mentioned in chapter 1, the NLTS2 direct assessment employed research versions of the Woodcock-Johnson III (WJ III) subtests for reading comprehensions, synonyms and antonyms, mathematics calculation, applied mathematics problem solving, and content knowledge in social studies and science. The research and published (i.e., standard) versions of the subtests share items and administration procedures. The difference between them lies in the larger number of items used in the standard version; the time (and, therefore, expense) of the standard version precluded its use for the large NLTS2 sample.

The research versions were created by the original test developers by reducing the item density from approximately three items per 10 W score points for the published version to one or two items for the research version, depending on the subtests. This is possible without changing the scoring or interpretation of the subtests because the WJ III is based on the Rasch model (Andrich 1988; Wright and Stone 1979), which allows for item-free measurement. Once the pool of items for a subtest is scaled per item response theory, different subsets that differ in item number and content can be used to create different versions of the test, with all subsets based on the same underlying Rasch-scaled measurement scale. Thus, the shorter research versions tests produce scores on the same scale as the full-length test and use the same national norms as those that underlie the published full-length tests. In addition to reducing the number of items, testing time also was reduced by changing the criteria for establishing basal and ceiling points from six consecutive correct items and six consecutive incorrect items, respectively, to three items.

Tests designed with these specifications have an average reliability of .65 and a standard error of measurement (SEM) of 10.0, in contrast to .85 and 5.7 for the publication-length tests. Although the individual SEMs are much larger for the research version, the important statistic for large-scale group analyses is the standard error of the mean, not the SEM, because the results are not used for individual programming decisions (e.g., eligibility for special education services). The standard error of the mean is a function of the standard deviation and sample size. Thus, if a test has a typical standard deviation of 15 W points, in a sample of 1,000, the standard error of the mean would be approximately 0.5, an acceptable level for calculating the group-level statistical estimates used in NLTS2.

Determining the Form of Assessment

Whether an age-eligible youth was administered a direct assessment or an adult was asked to complete a functional rating for him or her was determined through a screening process. For in-school youth, assessors conducted a telephone or in-person screening interview with the school staff person who was most familiar with a youth and his or her school program; in 91 percent of cases, this person was a special educator. Screening interviews were conducted with parents if youth were no longer in school. Screening information was used to determine whether a youth was able to participate in the direct assessment. To do so, a youth needed to be able to understand directions given in spoken English, large print, Braille, or sign language; have a consistent response mode (i.e., the assessor could reliably understand the youth's responses);[1] and the ability to work with an assessor or with someone who was familiar to the youth and who could and would conduct the assessment in the presence of the assessor.[2] If a youth met these criteria, additional questions were asked to determine which components of the assessment the youth could be administered.[3] The screening interview also sought to identify any accommodations that a youth required for the direct assessment. If a youth did not meet the requirements for the direct assessment, even with accommodations, he or she was deemed eligible for the functional rating.

Assessment Procedures

Direct Assessment

Hiring and training assessors. Assessors typically were school psychologists or teachers and were recruited in the geographic areas of eligible youth. Approximately 800 assessors were used in each wave of data collection, with the majority of 2002 assessors returning for the 2004 administration. Potential assessors submitted resumes and participated in a telephone interview to determine that they had experience conducting assessments of students with disabilities. The training of successful applicants consisted of reviewing the Field Assessors Guide, training

---

[1] School staff or parents were told, "The assessment requires that students (or youth) answer questions reliably," and then asked, "Is [YOUTH] able to reliably answer questions?"

[2] School staff or parents were asked, "Would [YOUTH] be able to answer questions asked by someone he/she doesn't know?" If the response was "no," the staff person was asked, "Would [YOUTH] be able to answer questions asked by someone he/she doesn't know if someone he/she knew was in the room?" If the response was "no," the person was asked, "Would [YOUTH] be able to answer questions if someone he/she knew asked the questions?" If the response was "yes," the person was asked, "Is there a person [YOUTH] knows available to conduct the assessment?" Across the two waves, 75 youth were reported to need a familiar adult to be present during or to administer the assessment in the presence of the assessment administrator. No statistically significant differences between youth reported to require such support and those not so reported were found on demographic factors; disability category; self-care, social, or functional cognitive skills; or mean standard scores on any assessment subtest. However, youth reported to need the presence of a familiar adult to complete the direct assessment were significantly more likely to have had their disabilities identified at birth than youth who were not reported to need this form of assistance (74 percent vs. 11 percent, $p < .01$) and were less likely to have had their disabilities identified at age 6 or older (7 percent vs. 37 percent, $p < .05$).

[3] To determine participation in the synonyms/antonyms and content knowledge subtests, school staff or parents were asked: "Can [YOUTH] read simple printed [or Braille] words, like 'road' or 'big'?" To determine participation in the mathematics calculation and applied problems subtests, the person was asked, "Can [YOUTH] recognize printed [or Braille] numbers?" To determine participation in the passage comprehension subtest, the person was asked, "Can [YOUTH] read written [or Braille] sentences?"

video, and the testing materials (WJ III direct assessment presentation "easel," test manual, and scoring booklets; functional rating scale; and screening interview questionnaire), and completing with 100 percent accuracy a self-administered test on the information and material presented in the Guide, video, and testing materials.

Each field assessor was assigned to a supervisor, who was available to answer questions about test administration, oversaw the training process, and reviewed and verified the successful completion of the field assessor test. When assessors successfully completed the training, they signed a work agreement and confidentiality pledge and were provided contact information, consent forms, and other assessment materials for the eligible youth in their area.

Conducting assessments. For youth who were able to participate in the direct assessment and who were still in school, the assessments generally were conducted at school when students were not in class. Some out-of-school youth also were assessed at the school they had once attended, but assessments for many out-of-school youth were conducted at youth's homes or in community settings.

Assessors contacted schools and parents to locate youth, identify a staff person who knew the youth well with whom to conduct the screening interview, and arrange for the appropriate assessment to be completed. The screening and direct assessment instruments (i.e., instructions and individual items) were fully scripted to maintain consistency of administration across assessors. Possible response choices for each item and instructions for scoring and for establishing basal and ceiling scores were included in the assessment easel, test manual, and scoring booklet. In the scoring booklet, assessors indicated only whether an item was answered correctly; all other scoring functions were conducted by NLTS2 project staff when booklets were submitted after completion of the assessment.

Use of accommodations. On the basis of recommendations of the assessment design panel and to be consistent with principles[4] underlying the inclusion of students with disabilities in standardized assessments (Thurlow, Quenemoen, Thompson, and Lehr 2001), the NLTS2 direct assessment procedure was designed to mirror students' day-to-day instruction and test participation with regard to the use of accommodations—i.e., a youth participating in the direct assessment was offered the same accommodations called for in his or her IEP for instruction and testing. The screening questionnaire requested information regarding a youth's need to take breaks during testing; use special furniture or lighting; have aides or assistants help with testing; or use American Sign Language (ASL), Braille, large print materials, or an abacus or calculator.

However, the design panel acknowledged that the nature of the WJ III as an untimed, individually administered test would make most accommodations used in state accountability testing (e.g., more time to complete the test) unnecessary. Consistent with this view, fewer accommodations were actually requested for youth in the assessment than they received in classes. Overall, 61 percent of youth received no accommodations, 28 percent received one accommodation, and 11 percent received two or more. The rate of receipt of specific accommodations is as follows: breaks (8 percent), special furniture or lighting (5 percent), an aide or assistant (5 percent), an ASL interpreter (8 percent), Braille (6 percent), and abacus or

[4] "Principle 3. All students with disabilities are included when student scores are publicly reported, in the same frequency and format as all other students, whether they participate with or without accommodations, or in an alternate assessment" (Thurlow, Quenemoen, Thompson, and Lehr 2001, p. 3).

calculator (23 percent). Those who participated in the direct assessment with one or more accommodations do not differ significantly from those who did not in disability-related factors, demographics, or mean standard scores on any direct assessment subtest.

It is important to note that norms for the WJ III were established for the general population, who would not need the accommodations provided to participants in the NLTS2 direct assessment. Thus, the NLTS2 procedures represent a departure from the standard WJ III procedures, but one that was deemed appropriate by the design panel for the population being assessed; without accommodation, some youth would have been unable to demonstrate competencies they in fact possessed, biasing downward measures of true achievement. The actual effects of providing a particular accommodation could be measured only by providing it to some students who required it and withholding it from others. However, analyses reported in chapter 4 attempt to estimate the relationship between provision of accommodations and academic achievement by including measures of their use in multivariate analyses, along with variables intended to control for variations in disability-related factors that could act as a proxy for need for such accommodations.

Functional Rating

If screening information indicated the direct assessment was inappropriate for a youth, a functional rating form and instructions for its completion were sent by the assessor to the youth's teacher if he or she was in school or to a parent if he or she was no longer in school or if the school would not participate in the assessment and rating process.[5] Assessors followed up with recipients to ensure an acceptable response rate. Completed rating forms were returned directly to NLTS2 project staff in postage-paid envelopes provided for that purpose. Respondents were compensated at the rate of $30 for each completed functional rating.

Analysis Approaches

Analyses reported in this document involve simple descriptive statistics (e.g., frequencies, means), correlational methods (i.e., cross-tabulations), and multivariate models (i.e., ordinary least squares regression). With the exception of seven variables that are included in the multivariate models, these analysis approaches exclude cases with missing values; imputation conducted for the seven exceptions is described in appendix A.

Regarding cross-tabulations, statistically significant differences between subgroups (e.g., youth in different disability categories) are identified using *F* tests. This approach has been followed because in all cases, the intent is to identify significant differences between two specific groups (e.g., youth with learning disabilities and those with mental retardation), rather than identifying a more general "disability effect" on the distribution of the variable of interest. In the case of unweighted data, comparing two percentages is usually accomplished using nonparametric statistics, such as the Fisher exact test. In the case of NLTS2, the data are weighted, and the usual nonparametric tests would yield significance levels that are too small,

[5] Approximately 22 percent of youth with a functional rating are estimated to have had it completed by a parent. Only one statistically significant difference between those with ratings that were parent-completed and those with ratings completed by teachers was noted on the variety of demographic and disability-related factors and assessment scores examined. The group with ratings completed by parents had a significantly larger proportion of African American youth than the group with teacher-completed ratings (43 percent vs. 14 percent, $p < .05$).

because the NLTS2 effective sample size is less than the nominal sample size. The p-values for the test statistic used as an alternative approach to determine statistical significance are derived from an *F*(1, infinity) distribution (i.e., a chi-square distribution with one degree of freedom).

Multiple linear regression techniques are used in this report to assess the independent relationships between ordinal measures of academic achievement and characteristics of individual youth, their households, and their school program and experiences.[6] NLTS2 multivariate analyses and correlations are unweighted. Results are reported for analyses that include the full set of individual, household, and school factors simultaneously. This approach allows the modeling of the simultaneous influence of predictor variables on the dependent variable and provide estimates of model fit.

Youth to Whom Findings Generalize

As noted in chapter 1, the universe to which the NLTS2 sample generalizes is a cohort of students who were ages 13 through 16 and received special education services in grade 7 or above in participating schools and school districts as of December 1, 2000. Weights for analyses reported in this document are calculated so that all youth with either a direct assessment or a functional rating, taken together, generalize to that cohort, without regard to when the assessment was done or which form of assessment was done.

To illustrate, consider the following groups:

A = The entire NLTS2 sample.

A1 = The portion of A who are ages 16 through 18 as of the Wave 1 assessment.[7]

A1a = The portion of A1 who would be able to participate in the direct assessment.

A1b = The portion of A1 for whom the functional rating is more appropriate to their abilities.

A2 = The portion of A who are ages 16 through 18 as of the Wave 2 assessment.

A2a = The portion of A1 who would be able to participate in the direct assessment.

A2b = The portion of A1 for whom the functional rating is more appropriate to their abilities.

For each of these sample groups, there is a corresponding group in the universe, which can be denoted with a "B," such that the universe is B, the portion of the universe that is 16 through 18 as of the Wave 1 assessment is denoted B1, the portion of B1 who would be able to participate in the direct assessment is denoted B1a, etc. The sizes of these universe subgroups can be estimated by weighting all youth in A (as if they all had been respondents) up to the entire

[6] Multiple linear regression equations involve a linear combination of a set of independent variables in the following algebraic form: $Y' = a + b_1X_1 + b_2X_2 + ... + b_nX_n$, where Y' is the predicted value of the dependent variable, a is the constant or Y intercept, bs are the partial regression coefficients, and Xs are the values of the independent variables.

[7] This group also includes 10 students who had recently become 19 at the time of their Wave 1 assessment.

universe, B. Then the sum of the weights of youth in A, A1, A1a, A1b, etc. are estimates of the number of youth in B, B1, B1a, B1b, etc.

However, not all students in A1a, A1b, etc. were respondents. Let respondents in each subgroup be denoted by appending an "r" to the label (e.g., A1ar, A1br, etc.). Then weights can be computed (adjusting for various youth and school characteristics used as stratifying or post-stratifying variables) that project A1ar up to B1a, A1br up to B1a, A2ar up to B2a, and A2br up to B2b, that is

- Youth who participated in the direct assessment in Wave 1 represent the portion of the universe who were 16 to 18 as of the Wave 1 assessment and would be able to participate in the direct assessment.
- Youth for whom a functional rating was completed in Wave 1 represent the portion of the universe who were 16 to 18 as of the Wave 1 assessment and whose abilities would make the functional rating appropriate.
- Youth who completed the direct assessment in Wave 2 represent the portion of the universe who were 16 to 18 as of the Wave 2 assessment and would be able to participate in the direct assessment.
- Youth for whom a functional rating was completed in Wave 2 represent the portion of the universe who were 16 to 18 as of the Wave 2 assessment and whose abilities would make the functional rating appropriate.

Additional technical information is presented in appendix A.