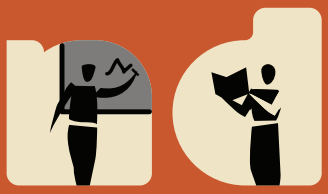
A BRIEF GUIDE TO SELECTING AND USING PRE-POST ASSESSMENTS

Prepared by the National Evaluation and Technical Assistance Center for the Education of Children and Youth who are Neglected, Delinquent, and At-Risk



neglected-delinquent.org

The content of this document does not necessarily reflect the views or policies of the U.S. Department of Education. This document was produced by NDTAC at the American Institutes for Research with funding from the Student Achievement and School Accountability Programs, OESE, U.S. Department of Education, under contract no. ED-04-CO-0025/0006. Permission is granted to reproduce this document.

Table of Contents

Summary To-Do List

This is a basic checklist for selecting and administering a pre-post test. Please refer to the remainder of this guide for additional detail on each of these elements.

- ❑ **Make sure the assessment is designed to be and is used as a pre-post test.**
  - Review the test publisher's materials. The test publisher should note that the assessment is appropriate for multiple administrations and has multiple, comparable forms available.
  - Verify that the publisher's recommendations for post-testing align with your assessment cycle (e.g., after a period of 90 days).
  - Review your testing procedures to make sure that you are giving different forms of the assessment at the pre- and post-test.

- ❑ **Verify that the test is appropriate for your students given their ages, skill levels, backgrounds, etc.**
  - Review the test publisher's materials. The publisher should note that the assessment is appropriate to administer to individuals at your students' age *and* skill level.
  - Verify that the publisher pilot-tested the assessment on students similar to your students (age, skill level, ethnicity, etc).
  - Ideally, the test should be adaptable to a variety of students and allow for "out of level" or "off grade-level" testing. This feature is most common in computer-adaptive tests.

- ❑ **Check that the test measures what you want it to measure.**
  - Review the test publisher's materials. The publisher should detail exactly what the assessment is designed to measure. If you want to assess reading ability, for example, as required for Federal reporting, a general intelligence test is not appropriate.
  - Require publishers to provide alignment studies of the test to State or National content standards; if not available, consider conducting a study internally.

- ❑ **Consider using the same assessment as your peers.**
  - Look at other programs, facilities or States to find out which assessments they are using. If your populations are similar (in size, age, ethnicity, etc.) consider adopting their assessments. Similar programs that are succeeding in increasing academic achievement can serve as a benchmark against which you can measure your progress.

- ❑ **Try to ensure that the conditions under which the student takes the pre- and post-tests are as similar as possible.**
  - Dedicate time and space within the school environment for assessments. This will help ensure that assessment conditions are consistent. Ideally the testing environment will be free from distractions, and students will have the amount of time specified by the publisher to complete the assessment. Be sure the person administering the assessment follows the exact same protocol each time, including how s/he answers students' questions.

A Brief Guide to Selecting and Using Pre-Post Assessments

Why this Guide?

This guide is for State, agency, and/or facility administrators who provide education for children and youth who are neglected, delinquent, or at risk (N or D). Primarily, it is intended as a resource for those who are in the process of choosing a new pre-post assessment or who wish to evaluate their current testing procedures.

The guide provides basic information about the ideal characteristics of a pre-post test and highlights important features to consider when requesting and evaluating information from test publishers. While few tests will have all the characteristics of an ideal pre-post test, this guide should help administrators select and use an instrument that best meets the needs of their student population.

What do we mean by "pre-post testing"?

"Pre-post testing" refers to academic achievement tests (in reading, math, and other subjects) that are given to students to assess their academic progress from the beginning to the end of a program of instruction. Applied to students who are N or D, the tests may be administered to students when they enter the facility and then again at exit from the facility. Both administrations should involve the use of the same test with alternative, comparable forms to ensure that the results from the administrations will be comparable and that the student is not taking the same test twice.

Ideally, the results of the pre-test reflect a student's abilities upon entry to an educational placement and provide a baseline for their current achievement level. Differences between the pre-test and post-test should reflect the learning that occurred while in the facility or program.

Pre-post testing does not have to be limited to administration at entry and upon exit. Because many students leave programs with little or no notice, there is not always time for post-testing before a student moves on to their next placement. Regular post-testing of academic achievement (e.g., every 90 days) ensures that students leave a facility with as complete an educational history as possible. However, post-testing should not exceed the number of unique forms available for the test being used and should follow the testing frequency guidelines outlined by the test publisher. Computer-adaptive tests are one possible option for programs conducting more frequent testing.

Why is pre-post testing critical for students who are N or D?

Pre-post testing is valuable to *students* because it illustrates and documents their individual academic gain. Timely pre-testing allows for accurate placement of the student upon entry to a new program, so the student can get started on a curriculum shortly after entering the facility, and consistent post-testing helps to ensure that students leave a facility with updated educational records and receive credit for their progress. In addition, being able to show students the

progress they have made while in the program can be a powerful tool to keep them invested in their education.

Pre-post testing is valuable to *teachers* because it allows for real-time progress monitoring. Because multiple post-tests can be administered throughout a student's enrollment, educational gains can be monitored and instruction can be adjusted appropriately. Measuring academic progress through appropriately administered pre-post tests can be a powerful tool in providing teachers feedback about how to better meet students' academic needs.

Pre-post testing is valuable to *facilities, programs, and State and local administrators* because the aggregated results of students' academic progress can be used for evaluating and improving educational programming. Collecting data on student progress allows administrators to measure the effectiveness of their educational programs and make any necessary changes, and provides local and State administrators with the necessary information to make comparisons across programs. In doing so, administrators can identify particularly effective programs that can serve as models for others. Such data collection can also help programs, agencies, and States meet State and Federal reporting requirements.

Having updated academic achievement data can also *improve and facilitate the transition process for* facilities, programs, schools, and agencies when students move between placements or return to district schools. Data from pre-post tests allow administrators to easily assess a student's skill level, provide credit, and place the student accordingly.

Clearly, there are a multitude of advantages to administering pre-post assessments. However, in order to maximize these benefits, it is essential that the assessment selected be appropriate both for pre-post testing and for the specific population.

Why is annual, State NCLB-required testing not enough?

State assessments are generally not appropriate tools for N or D programs to use to demonstrate the progress of their students for several reasons:

- Scores cannot be used to see individual student gain; scores are examined to see how all students at one grade perform compared to the same grade in another year.
- The State assessment is usually not designed to be given more than once per year; hence, it is not an appropriate pre-post test.
- Most of the students will have entered and exited the N or D system before the one-year mark (only about 10 percent of students are in facilities for more than a year[1]) and may not be enrolled during the administration of the Statewide test.

While the results of State assessments are useful in many other ways, they are not intended or designed to measure the academic progress of specific students *within* a given school year or reporting year.

[1] Snyder, H. & Sickmund, M. *Juvenile Offenders and Victims: 1999 National Report*, p. 201. Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention, 1999.
Data Source: Office of Juvenile Justice and Delinquency Prevention. *Census of Juveniles in Residential Placement 1997* [machine-readable data file]. Washington, D.C.: OJJDP, 1998.

What do the terms "reliability" and "validity" mean, and why are they important in assessment?

Reliability refers to the repeatability of a given testing instrument: the extent to which a student would be expected to perform similarly across multiple administrations of the test under similar conditions.

For example, you would expect that if you took the temperature outside one minute and then did it again the next minute, the thermometer would give the same (or a very similar) reading. The thermometer is a reliable instrument. In assessment, a test given to a student today should give approximately the same score as a [different version of the same] test given to that same child tomorrow, assuming that no academic program has taken place in the intervening day that should affect the student's performance.

Validity refers to the extent to which a given test is measuring what it is expected to measure.

For example, if a student gets a question wrong on a math test because s/he could not read and understand the word problem (a reading difficulty issue), but s/he could do the math in the question, then the question is also measuring the student's reading ability, not just his/her math ability. In this case, the question may not be a valid indicator of the student's math ability.

Outlined in the boxes are the types of reliability and validity to take into consideration when both selecting and administering a pre-post test. This information also highlights the types of information to request from a test publisher when comparing testing instruments.

Reliability

1. ***Test-retest reliability.*** Test-retest is a method for assessing the reliability of an instrument by administering the same test at different points in time. The correlation between the scores across the two test administrations should be high.

   Does the publisher of your test describe how they checked for test-retest reliability?

2. ***Alternate forms reliability.*** Alternate forms reliability refers to the relationship between scores on alternative forms of the same test, administered closely in time. In the case of pre-post testing, alternate forms reliability is more important than test-retest. When administering a test with multiple forms, you want to be sure the change in scores on the pre- and post-tests reflect students' abilities and not differences in the difficulty of the forms. Publishers with more than one test form should provide information about the similarity of student performance across the multiple forms.

   Does the publisher of your test describe how they checked for the reliability of alternate forms?

Validity

1. ***Construct validity*** refers to evidence that a test is measuring the content and skills (or "construct") it claims to measure. Although there are many types of evidence that fall under this category, two common types are *convergent validity* and *discriminant validity*.

    a. Convergent validity refers to the extent to which the scores on a given test are related to scores on other tests designed to measure the same, or very similar, constructs (e.g., students' scores on a test designed to measure reading comprehension should be highly related to scores from other tests designed for the same purpose).

    Does the publisher of your test provide evidence that performance on that test is positively related to other tests in the same content area?

    b. Discriminant validity, on the other hand, refers to the extent to which the scores on a given test are related to scores on tests designed to measure something different. In particular, scores from tests measuring different skills or abilities should be less highly correlated than tests measuring the same or similar skills or abilities.

    Does the publisher of your (reading) test provide evidence that performance on that test is less positively related to tests of other subjects (such as math) than it is to other tests of the same subject (i.e., reading)?

2. ***Content validity*** is the extent to which the items/tasks on a given instrument adequately *cover* the area of interest. For example, in a test of reading comprehension, all of the items/tasks on the instrument should: a) align with the skills associated with reading comprehension, and b) represent the *full range* of skills associated with reading comprehension. Skills *not* associated with this content area, such as knowledge of science or social studies, should not influence performance on the test.

    Does your test publisher describe how they checked for content validity?

3. ***External validity*** refers to the extent to which the results of a test given to one population of students are relevant to other groups of students. For example, test norms based on data collected from students enrolled in traditional classrooms may not be as useful for gauging the progress of students in alternative types of educational programs, especially if the students in these programs differ with respect to factors such as socioeconomic status, age, etc.

    Was your current test pilot tested and normed with students who match the student population in your facility or program (in areas such as ethnicity, age, academic skills, etc.)?

What are the four essential characteristics of a pre-post test?

Few tests exist that are perfectly valid, can be generalized to all populations, and assess students in all of the areas of interest. However, there are some critical features of pre-post tests that should not be ignored when selecting an instrument for N or D facilities and programs. These features and related questions are highlighted here.

1) The Test Should Be Designed for Multiple Administrations

The tests you choose should be designed to be taken repeatedly within a one-year period. Tests that were originally designed for administration on an annual basis are not appropriate, as they may not be sensitive to small changes in students' skill levels.

** Is your test designed to be administered more than once per year? If yes, how frequently?*
** Are you administering a test more often than is recommended by the publisher?*

2) The Test Must Have Multiple Forms
The test should have *at least* two different versions available for administration at the pre- and the post-test. Administrators should not post-test a student with the same questions they encountered in the pre-test. Doing so can produce invalid data because a student's progress cannot necessarily be attributed to the skills they have developed if they are already familiar with the test questions.

For example, students who are given the same reading comprehension exam during a short period of time may be able to remember the passage and then be able to skim through it more quickly. As a result, the student may be able to fill out more questions on the overall test and subsequently obtain a higher score. Such an increase would not be attributable to an improvement in skills, but rather to the student's familiarity with the test.

** Does your test have multiple forms so that you are never giving the same test to the same student?*
** If computer-administered and interactive, does your test have a sufficiently large question (or "item") bank so that the likelihood of a student receiving the same questions across test administrations is minimal?*

3*) Test Content Should Match the Skills of Interest*
The test chosen should be designed to measure the specific content areas of interest. Programs may have an interest in assessing progress in a number of areas: reading, writing, developmental skills, etc. For Federal reporting purposes, N or D programs currently are required to assess student progress in reading and math. Many tests are available that assess multiple content areas, or individual tests can be selected in each content area.

** What is the academic area you are interested in assessing?*
** Does the test you have selected align with and cover this content area?*

Below are examples of tests that are *not* appropriate for pre-post testing in reading and math:

- Writing and language assessments should not be used as reading assessments.
- Intelligence or IQ tests are not appropriate measures of student academic progress. Scores on these tests are not reflective of specific academic skills, and students would not be expected to demonstrate improvement during enrollment, as cognitive abilities do not typically change over a short period of time.

4) Test Design Should Match the Appropriate Student Population
While there are many options available for pre-post tests, the test selected should be relevant to the students served. Tests are often designed to be used with a specific age or grade range. The test administered should capture the actual skill levels of the students in the N or D program.

** What is the predominant age range and academic skill level of the students in your facility?*
** Are academic skill levels generally below the students' age group?*
** Does the test you have selected align appropriately with the skill levels of your students?*
** Does the test publisher provide evidence that the test is normed with neglected, delinquent, or other alternative student populations?*
**Is the test able to accommodate special needs, or is a different test necessary for students with disabilities?*

These questions are important for multiple reasons:

- Tests designed for high school students are not appropriate if the program population ranges from 9-21 years of age.
- Many students enter N or D programs testing "out of level" and usually lower than their age-equivalent peers. Programs may want to choose tests based on predominant skill levels rather than focusing on the age range of the students. Ideally, the test selected should either 1) provide the ability to test students who are off level, or 2) be selected to target the most common skill range of those enrolled.
- Many students who are N or D also have learning, behavioral, mental, and/or physical disabilities. Alternative testing options, as outlined by the Individuals with Disabilities Education Act (IDEA),[2] may need to be made available.

What other test characteristics should be considered?

In addition to the information provided above, there are several other features you may want to take into consideration when selecting a test. These include the type of test and the manner in which the test is administered.

1) Type of Test
The two most common types of tests are *criterion-referenced tests* and *norm-referenced tests*. The type of test selected should be based on the information you are most interested in knowing about your students and the type of information you are required to report. Pre-post tests are available in both formats.

> *Criterion-referenced tests* are geared toward examining student achievement in relation to a level of performance that has been defined in advance. Scores from criterion-referenced tests are often provided in relation to "achievement" or "proficiency" levels, which are basically evaluative statements defining the skills a student has obtained. Achievement levels can be prescriptive or descriptive. Proficiency levels are prescriptive when they specifically define what a student should and should not be able to do at a given grade. In contrast, descriptive achievement levels are not linked to specific grades or ages. Rather, they are concerned with student growth along a continuum, and provide

[2] N or D programs should be knowledgeable about the requirements of the Individuals with Disabilities Education Act (IDEA) as they apply to youth who are neglected or delinquent. Information specific to the IDEA 2004 Legislation can be found at http://www.ed.gov/policy/speced/guid/idea/idea2004.html

a mechanism for monitoring how a student is progressing. Criterion-referenced tests may provide more information than norm-referenced tests when tracking student progress.

Norm-referenced tests provide information about how a particular student's performance compares to that of other students who have taken the test. The test score itself does not have meaning without information describing the students in the comparison group. Two common types of scores generated from norm-referenced tests are grade equivalent scores and percentile scores.

- Grade equivalent scores are created by obtaining raw scores from a large sample of students and then using those scores to develop a scale demonstrating the average scores of students at specific grade or age levels. These scores are often broken down into increments of expected performance by age or grade.

- *Percentile scores* provide a number indicating how a student did in comparison to other students who took the test. For example, if a student has a percentile score of 70, this means that 70 percent of students who took the test received the same score or lower, and 30 percent of the students received a higher score. A student's percentile can change depending on the population of students to which the student is being compared. For example, the ranking could be based on (a) the sample from which the test was normed, or (b) the other students who took the test within a given time period (such as students in the State, district, or school). When using a percentile score, it is important to be aware of the population on which the scores are developed. This is especially true for the population of students who are N or D, as they are not easily compared with their same-aged peers.

In summary, the main difference between these two types of tests is that criterion-referenced tests compare a student to a predetermined set of standards (regardless of how other students perform), while norm referencing compares a student's performance to that of other students who took the same test. However, note that these two categories of tests are not mutually exclusive. In fact, there are a number of tests for which both criterion- and norm-referenced information is available.

2) Format of Test Administration

The format of test administration refers to both the manner in which a student is presented with the test questions, and the manner in which they are required to respond to the questions. While this may seem to be a minor issue, the format in which a test is administered can impact a student's test scores.

For instance, a student who cannot read cannot take a written exam effectively. However, the student may be able to demonstrate other skills if the test is given orally. Similarly, computer-based administration may not be appropriate if most of the students in the program cannot type or do not know how to use a computer (see also mode effects in the section below). Knowing the average test-taking skills of the students in a facility or program will help determine which type of administration is most appropriate.

**Are the majority of students in my educational program…*
…able to read and write?
…fluent in English?
…familiar with how to use computers?

Oral administration refers to both the format in which the test is given and the format in which students respond. A teacher or administrator may ask the questions aloud, while the students may either respond aloud or write the answers. Students who respond aloud would most likely be given the test in a one-on-one setting rather than in a group. Oral administration also requires that trained staff be available for administering the tests.

Group vs. individual testing refers to whether students are given the test in a large classroom setting all at the same time, or whether they take the test alone or one on one with a teacher. Because students in N or D programs are often entering and exiting the facilities at various times during the year, individual administration may be more appropriate for pre- and post-testing.

Pencil and paper tests are often read alone by the individual student, though they could be given orally. The responses are then written down by hand by the student. This is a very common testing format.

Computer-administered tests are tests that are given primarily by a computer. The computer presents the test to the student, and then the student is able to respond directly on the computer as well. A particular type of computer-administered test is referred to as *computer-adaptive tests,* which adjust the difficulty of subsequent questions based on the student's response (correct or incorrect) to those questions already administered.

Computer-based assessments have benefits worth considering. These include the ease of administration for the teacher or faculty (as students can take the tests individually and largely unsupervised), the ability to obtain scoring information immediately, and (if the test is administered adaptively) the potential ability to administer a wide variety of unique items during post-testing. On the other hand, potential concerns related to the implementation of computer-based testing include the availability of funding to invest in and maintain the technology, the ability of students in the program to effectively use the technology, and (if the test is adaptive) the capacity of the testing program's item bank to ensure that students will be given non-overlapping sets of items across pre- and post-test administrations.

Additional factors that may impact pre-post test results

Testing or practice effects. Students who take the same test repeatedly often show improvement on subsequent tests that can be attributed to familiarity with the questions or the testing format. This is an especially important factor in the pre-post testing environment. One solution is to make sure the test chosen has multiple (equivalent) forms available for post-testing. Another solution may be the use of computer-adaptive tests that often provide a large question bank.

Having a large pool of questions to choose from reduces the likelihood that students will receive the same questions that appeared in the first administration.

Fatigue effects. Fatigue refers primarily to factors that may occur during the test administration—such as students becoming hungry or tired over the course of the test. Such factors may have a negative impact on students' scores. One way to reduce the effects of fatigue is to make sure that the length of the test is not unreasonably long, or to allow for breaks during the testing period (if this is acceptable according to the test administration manual).

Motivational effects. A student's desire to do well on a test may also influence scores. For instance, a student may enter a facility with a higher skill level but not care about the pre-test upon arrival; then, for any number of reasons, the same student may become motivated to achieve and do well while enrolled. In such a case, the increase in the student's post-test scores may not be solely attributable to learning, as it may in part reflect the already existing skills that were not demonstrated during the pre-test. Conversely, a student who is preparing to leave a facility may not be motivated to put any effort into their exit exam, thus demonstrating less gain (or even a decline) than what actually occurred. Encouraging and supporting a student's desire to always do his or her best may help to minimize this effect.

Characteristics of the test administration situation. The manner in which a test is administered is another factor that can impact a student's results. For example, in a noisy and unruly classroom, a group-administered test may not be appropriate because distraction may prevent students from being able to focus. Similarly, a computer-administered test may not be appropriate if students are unfamiliar with the technology.

Length of test. It is hard to estimate the ideal length of a test. Tests that are too long may create boredom and cause students to lose focus. Conversely, tests that are too short may not adequately cover the area of interest and thus produce invalid scores. Administrators should choose tests based on what they know about the population of students they serve. For example, shorter tests may be more appropriate for younger students, students with lower skill levels, or students with behavioral or learning disorders.

What if I can't find a test that fits all of these criteria?

Currently, there are few tests that have been designed specifically for students who are N or D. However, lack of a perfect test does not mean that there are not tests out there that are better than others. Each facility or program should take all of these assessment criteria into consideration in conjunction with the needs of their students. These considerations should help programs make informed decisions regarding which assessment instruments and testing procedures to choose and implement.

Programs should revisit their testing procedures and re-examine the pool of available tests every few years to assess whether new tests exist, or if the student population has changed in a way that might warrant selection of a more appropriate test instrument. Furthermore, programs may want to consider paying for tests to be developed, or lobbying within their State for support in developing a uniform N or D assessment tool that can be used to measure student progress.

Related Information

Test selection and resources

The U.S. Department of Education does not restrict the type of tests used by Title I Part D programs to assess student progress in N or D programs, nor does it endorse the use of any specific test. However, the Department does expect the tests chosen by programs to be appropriate and valid for pre- and post-testing students who are N or D, and to be able to yield valid reporting data.

Based on feedback received from States, the National Evaluation and Technical Assistance Center for the Education of Children and Youth Who Are Neglected, Delinquent, or At Risk (NDTAC) is in the process of compiling information about the various assessments currently used in the field. When prepared, this information will available on the NDTAC Web site at www.neglected-delinquent.org.

The benefits of a Statewide assessment

Currently, facilities and programs throughout most States are using a variety of pre-post testing instruments. Such variety makes it very difficult to compile data and make comparisons across programs or districts since scoring systems can be highly disparate.

There are two ways to help alleviate this problem: 1) States can require that all N or D programs use the same pre-post assessment, or 2) States can recommend a subset of tests that are aligned with State reporting criteria, from which programs may choose. While these activities are not required, States may want to consider these options for streamlining the testing of students who are N or D. Doing so will allow States to collect data that can generate program comparisons and help indicate where programs are succeeding and where they may need additional assistance.

Title I Part D legislation related to academic achievement

In addition to State and local requirements, Federal regulations also require programs that receive funding for students who are neglected and delinquent to measure the academic progress of these students.

Section 1431 (Program Evaluations) States:
(a) Each State agency or local education agency that conducts a program under subpart 1 or 2 shall evaluate the program, disaggregating the data on participation by gender, race, ethnicity, and age, not less than once every 3 years, to determine the program's impact on the ability of students:

(1) to maintain and improve educational achievement;
(2) to accrue school credits that meet State requirements for grade promotion and secondary school graduation;
(3) to make the transition to a regular program or other education program operated by a local educational agency;

(4) to complete secondary school (or secondary school equivalency requirements) and obtain employment after leaving the correctional facility or institution for neglected and delinquent children and youth; and
(5) as appropriate, to participate in postsecondary education and job training programs.

(b) Exception: The disaggregation required under subsection (a) shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student.

(c) In conducting each evaluation under subsection (a), a SA or LEA shall use multiple and appropriate measures of student progress.

Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th ed.). New York: Harper & Row.

Feldt, L.S., & Brennan, R.L. (1993). Reliability. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Hopkins, K.D. (1998). *Educational and Psychological Measurement and Evaluation* (8th ed.). Boston, MA: Allyn & Bacon.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Peterson, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.