

Using Augmented Norm-Referenced Assessments for NCLB Compliance

Sasha Zucker

With assistance from:

Ray Christensen

Roy T. Ellis

Herb Harris

Duane Manning

August 2004



Using Augmented Norm-Referenced Assessments for NCLB Compliance

Introduction

Since the introduction of the *Stanford Achievement Test* in 1923—the first of its kind—large-scale standardized norm-referenced assessments have served a variety of important purposes in education. Roles that policymakers expect testing programs to serve include certifying a student’s level of achievement, providing information about an education system’s effectiveness, motivating student performance, bringing coherence to a curriculum, and holding schools and educators accountable for student performance (Hamilton, Stecher, and Klein, 2002).

With the advent of the accountability and standards-based reform movements of the 1980s, expectations grew for the information that tests would yield. These expectations led to an increase in the number and variety of specialized testing programs administered to students. However, observers have noted that the benefits derived from increased testing have been balanced, if not outweighed, by the reduction in instruction time available to students and the monetary expense of administering multiple testing programs (Linn and Hambleton, 1991).

The passage into law of the *No Child Left Behind Act* of 2001 (NCLB)—the latest reauthorization of the *Elementary and Secondary Education Act* of 1965 (ESEA)—has affected the recent development and use of educational assessments. NCLB requires states to implement rigorous annual testing programs in reading and mathematics for students in grades 3 through 8 and in one high school grade by the 2005–2006 school year. Also, by the 2007–2008 school year, students must be assessed in science at least once in grades 3 through 5, once in grades 6 through 9, and once in grades 10 through 12 (NCLB, § 6311). The main purpose of the tests mandated under NCLB is to determine the proportion of students proficient in each subject area with the goal of all students reaching proficiency by 2014.



Using Augmented Norm-Referenced Assessments

Many states are now adopting assessment systems designed to satisfy the accountability mandate of NCLB while also gathering the data that stakeholders (parents, students, and educators) need to evaluate and improve an education system. This report examines a new type of assessment that accomplishes this complex objective—the *augmented norm-referenced test*.

Traditional Standardized Assessment Designs

The design of a standardized educational assessment is determined largely by how its results are interpreted and used (Nitko, 2004). Depending on which type of interpretation is sought, test publishers use different types of research studies and development methods to design and produce the assessment. Traditionally, there are two types of standardized educational assessments: *criterion-referenced assessments*, which report a student's results in comparison to a curriculum, and *norm-referenced assessments*, which report a student's results in comparison to a group of other students who have taken the same assessment (Linn and Hambleton, 1991).

Criterion-Referenced Assessments

Criterion-referenced assessments—also called standards-based, standards-referenced (Hamilton et al., 2002), or curriculum-specific assessments (Linn and Hambleton, 1991)—are designed to measure what students know and can do in comparison to academic standards for a subject area (Nitko, 2004). To produce a measure that is useful to a state or local education agency, a criterion-referenced assessment must use items that match the content standards of the corresponding state or local curriculum (Linn and Hambleton, 1991). The measures most frequently associated with criterion-referenced assessments are *performance levels*, with familiar examples including basic, proficient, and advanced. A student's performance level is determined by comparing the student's score to the assessment's *cut scores*—thresholds between performance levels that are established during the assessment's development. Scoring at or above a cut score indicates that the student has reached that performance level. For example, if the cut score for an assessment's proficient performance level is set at 350 and the cut score for the advanced performance level is set at 450, a student who scores between 350 and 449 has reached the proficient level and a student who scores 450 or higher has reached the advanced level.

While criterion-referenced tests are useful for evaluating students according to performance levels, they are less effective at obtaining other kinds of information. Although criterion-referenced test results may indicate whether a student has reached a higher performance level, there is no distinction between students who are slightly below a cut score and students who are far from reaching the cut score



Using Augmented Norm-Referenced Assessments

(Hamilton et al., 2002). Hence, criterion-referenced assessments lack the specificity to identify the relative strengths and weaknesses of students or their progress in a particular subject area—two indicators that districts and schools often find useful for monitoring and improving the performance of individual students and education programs.

Norm-Referenced Assessments

To determine the strengths and weaknesses of a student in a subject area, education agencies rely on the administration of norm-referenced assessments, such as the *Stanford Achievement Test Series, Tenth Edition* (Stanford 10). The results from a norm-referenced test compare a student's achievement in a subject area to a nationally representative sample of students, referred to as the *norm group*. To make this comparison possible and the results valid, the test is developed using rigorous scientific research programs in which it is administered under standardized conditions to the norm group. In subsequent administrations of the published assessment, the results of individual students are compared to the results of the norm group using measures that rank-order each student, such as percentile ranks or stanines (Nitko, 2004).

States rely on published norm-referenced tests, such as Stanford 10, to gather information about their student populations and to compare the quality of their educational systems to those of other states (Hamilton et al., 2002). Because of their national scope, norm-referenced assessments typically measure a sample of the academic content taught in schools nationwide. Test items are developed by consulting with educators, examining current textbooks, and reviewing state content standards (Linn and Hambleton, 1991). With the advent of state content standards, publishers have been able to develop norm-referenced assessments that measure the content standards which states hold in common.

NCLB Requirements for Standards and Assessment

NCLB requires each state to adopt challenging academic content and achievement standards for all public school students and to implement a set of high-quality, yearly student academic assessments that measure these standards. Moreover, to provide meaningful information about student achievement, the assessments must report the student's results using at least three performance levels—basic, proficient, and advanced (NCLB, §6311). States may choose different names for their performance levels or may use more than three performance levels. For example, Louisiana designates five performance levels with the names unsatisfactory, approaching basic, basic, mastery (proficient), and advanced (Louisiana Department of Education, 2000; U.S. Department of Education, 2003).

Using Augmented Norm-Referenced Assessments

Allowance for Augmented Norm-Referenced Assessments

At first glance, the assessment requirements found in NCLB describe the qualities of criterion-referenced assessments. One might conclude that, to comply with NCLB, states are limited to the use of criterion-referenced assessments developed specifically to match their content standards. However, this apparent limitation raises a concern that states will reduce the use of norm-referenced tests upon which stakeholders rely to evaluate and improve their education systems (Hamilton et al., 2002).

Following the passage of NCLB into law, a resolution was sought for the perceived conflict between the apparent mandate for criterion-referenced assessments and the continued need for the data provided by norm-referenced assessments (Hamilton et al., 2002). In formulating the rules for NCLB, the U.S. Department of Education allowed for a solution: norm-referenced tests that have been *augmented* to align with a state's standards. At the 2003 annual conference of the Association for Supervision and Curriculum Development, Associate Deputy Undersecretary Tom Corwin of the U.S. Department of Education discussed this approach:

The question facing the Department, as we began to implement [NCLB], was whether the Act allows the use of norm-referenced tests, and whether some use of NRTs [norm-referenced tests] . . . would be good practice within the overall framework of the statute. We came to this issue after finding, in our implementation of the previous reauthorization [of ESEA], that some states could augment NRTs to include more test items aligned with state standards . . . in a manner that assessment experts believed could meet the test of alignment with state standards and meet the other requirements of the law—requirements in such areas as test quality, validity, and reliability, and a test's use of multiple measures for assessing student achievement.

We thus allow, in our regulations, states to operate assessment systems that combine both criterion- and norm-referenced tests, or include augmented NRTs alone, so long as . . . strict standards of quality and alignment are met. . . . And, clearly, allowing the use of norm-referenced tests, as a part of a broader system or if carefully augmented, has facilitated implementation of the law in a number of states. (Corwin, 2003)

In accordance with this discussion, the latest draft of the U.S. Department of Education's *Standards and Assessments: Non-Regulatory Guidance* (2003) for NCLB makes the following allowance for augmented norm-referenced tests:

Using Augmented Norm-Referenced Assessments

A state may include either criterion-referenced assessments or augmented norm-referenced assessments in its assessment system. States wanting to use a norm-referenced assessment at a particular grade must augment that assessment with additional items as necessary to accurately measure the depth and breadth of the state’s academic content standards, and the assessment must express student results in terms of the state’s student academic achievement standards. (pp. 13–14)

An augmented norm-referenced test is further defined by the U.S. Department of Education’s *Standards and Assessments: Non-Regulatory Guidance* (2003, p. 14) as “an assessment in which selected items from a norm-referenced assessment are combined with additional items written specifically to assess state content standards not covered by a norm-referenced assessment.” To augment a norm-referenced test, a state must carry out an independent alignment study to determine the degree of the match between the content and depth of the standards and the items in the assessment (Ananda, 2003). There are several rigorous, widely accepted approaches for performing alignment studies; Pearson Education, Inc. (Pearson) frequently relies on a method developed by Norman Webb (1999). Once the gap in content and depth has been identified, the test developer can achieve the match by adding items to the norm-referenced assessment. It is worth noting that alignment to academic content standards is required for any assessment used to satisfy NCLB requirements, including criterion-referenced assessments, and states must provide evidence of the alignment process (U.S. Department of Education, 2003).

Scientific Background of Augmented Norm-Referenced Assessments

Encouraging the use of augmented norm-referenced assessments represents an “evolutionary rather than revolutionary” aspect of NCLB that is supported by leading educational assessment research and practices (Corwin, 2003). With the increased number of testing programs introduced during the education reforms of recent decades, educators have been seeking to reduce the time and cost of testing while also yielding more data from the tests administered (Linn and Hambleton, 1991). Faced with the dilemma of the need for both criterion-referenced and norm-referenced information, educational researchers began investigating the possibility of combining the two designs. The main concern of the researchers was the validity of scores derived from this new type of assessment (Linn and Hambleton, 1991).

In a series of studies during the 1980s, education researchers investigated different models for combining norm-referenced and criterion-referenced tests, including the augmented norm-referenced model already described (Linn and



Using Augmented Norm-Referenced Assessments

Hambleton, 1991). Another model investigated, sometimes called a “hybrid” assessment, uses a criterion-referenced assessment as its base and obtains the norm-referenced measure of the student’s achievement by concurrently administering a subset of norm-referenced items in the same or in a separate booklet (Skinner and Staresina, 2004). In both models, norm-referenced item content is augmented to match the academic content standards for the subject area (Linn and Hambleton, 1991).

The studies of the 1980s primarily investigated the effects that these models had on the validity of the results. Researchers investigated the correspondence between augmented norm-referenced assessments and the unmodified versions, the impact of the increased length of the customized assessment on student examinees, and the effect of administering norm-referenced items adjacent to criterion-referenced items—a phenomenon known as context effect. The results of these studies demonstrated that carefully developed augmented norm-referenced assessments can yield valid criterion-referenced and norm-referenced measures of student achievement (Linn and Hambleton, 1991).

Since these studies have been completed, the augmentation of commercially developed norm-referenced tests has been recognized as an innovative and important solution for states faced with NCLB compliance. A report by the Education Leaders Council (AccountabilityWorks, 2002) cites the use of customized off-the-shelf norm-referenced assessments as an innovative solution that is less expensive and faster to develop than assessments that are custom-developed for a state. Moreover, this solution can be implemented immediately by states seeking to comply with NCLB.

Experts in the field of educational assessment cite augmented norm-referenced tests as a compelling solution to NCLB’s assessment mandate. In an interview with the National Governor’s Association (2002), Robert Linn, co-director of the National Center for Research on Evaluation, Standards, and Student Testing, asserts that “for states that do not already have in place all the assessments required by NCLB, norm-referenced tests that are augmented to provide adequate coverage of state content standards in grades and subjects where state assessments are not in place will often be the most efficient and cost-effective way to fill in missing grade/subject areas.” The combination of these expert opinions with the scientific education research of the 1980s supports the use of augmented norm-referenced assessments as a valuable new type of instrument for education.

Augmented Norm-Referenced Assessments in Use Today

The combination of supporting research, practical success, and NCLB’s allowance for augmented norm-referenced assessments has led more states to

Using Augmented Norm-Referenced Assessments

adopt this solution for their statewide testing programs. A recent report lists 12 states that are currently using augmented norm-referenced tests to comply with the NCLB accountability mandate (Skinner and Staesina, 2004). Several of these states—Alabama, Delaware, Hawaii, Maryland, New Mexico, and South Dakota—use an augmented version of the *Stanford Achievement Test Series*, Ninth Edition (Stanford 9) or Tenth Edition (Stanford 10) to obtain both norm-referenced and criterion-referenced scores from their testing programs. To augment Stanford 9 or Stanford 10, Pearson’s assessment specialists work with state officials and educators to perform an alignment study that identifies the existing items that match state standards. Item writers from both the state and Pearson use the results of the study to develop new items for the assessment so that it can produce both criterion-referenced and norm-referenced scores.

Conclusion

By augmenting norm-referenced assessments, educational agencies are able to use high-quality, relatively inexpensive standardized assessments to satisfy the accountability mandate of NCLB. Stanford 10, the nation’s leading norm-referenced academic assessment with state-of-the-art features—including untimed administration, full-color materials, and universal design—engages students and provides them with the best opportunity to show what they know and can do. Augmenting Stanford 10 to match academic content standards empowers educational agencies to satisfy NCLB assessment requirements and to obtain useful norm-referenced data about their student populations. Pearson’s assessment specialists have demonstrated competence in performing alignment studies and in custom-developing items to fill any content gaps. By collaborating with Pearson to produce valid and effective assessment systems, education agencies can meet accountability mandates while also obtaining information needed to continuously improve education for future generations of students.

References

- AccountabilityWorks. (2002). *State innovation priorities for state testing programs*. Washington, DC: Education Leadership Council. Retrieved from <http://www.educationleaders.org/elc/events/st%20innovation.pdf> on April 2, 2004.
- Ananda, S. (2003). *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd.

Using Augmented Norm-Referenced Assessments

- Corwin, T. (2003, March 21). *The No Child Left Behind act: Where are we now? Where are we going?* Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/print/news/speeches/2003/03/03082003.html> on April 7, 2004.
- Hamilton, L. S., Stecher, B. M., & Klein, S. P. (Eds.). (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: Rand Corporation.
- Linn, R. L., & Hambleton, R. K. (1991). *Customized tests and customized norms*. Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Testing.
- Louisiana Department of Education. (2000). *LEAP 21 achievement levels*. Baton Rouge, LA: Author. Retrieved from <http://www.doe.state.la.us/lde/uploads/1244.pdf> on August 9, 2004.
- National Governors Association. (2002, July 26). *NCLB: Interview with Robert Linn about assessment*. Retrieved from http://www.nga.org/center/divisions/1,1188,C_ISSUE_BRIEF%5ED_4154,00.html on February 12, 2004.
- Nitko, A. J. (2004). *Educational assessments of students*. Englewood Cliffs, NJ: Prentice Hall.
- No Child Left Behind Act of 2001, 20 U.S.C § 6311 et seq.* (2001).
- Skinner, R. A. & Staesina, L. N. (2004, January 8). State of the states. *Education Week 23 (17)*, 97-99.
- U.S. Department of Education. (2003). *Standards and assessments: Non-regulatory guidance*. Washington, DC: Author.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Washington, DC: Council of Chief State School Officers.

Additional copies of this and related documents are available from:
Pearson Education, Inc.
19500 Bulverde Road
San Antonio, TX 78259
1-800-211-8378
1-877-576-1816 (fax)
<http://www.pearsonassess.com>