Recommendations for Building a Valid Benchmark Assessment System: Second Report to the Jackson Public Schools

CRESST Report 724

David Niemi, Jia Wang, Haiwen Wang,
Julia Vallone, and Noelle Griffin
CRESST/University of California, Los Angeles

July 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Stupski Foundation's Support of Assessment Development: Jackson Public Schools program, Award #: 20050546. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the Stupski Foundation.

RECOMMENDATIONS FOR BUILDING A VALID BENCHMARK ASSESSMENT SYSTEM: SECOND REPORT TO THE JACKSON PUBLIC SCHOOLS

David Niemi, Jia Wang, Haiwen Wang, Julia Vallone, and Noelle Griffin

CRESST/University of California, Los Angeles

Abstract

There are usually many testing activities going on in a school, with different tests serving different purposes, thus organization and planning are key in creating an efficient system in assessing the most important educational objectives. In the ideal case, an assessment system will be able to inform on student learning, instruction and curricula, and district and school administration, as well as providing information to identify and solve educational problems. This report represents the second of two deliverables provided to the Jackson Public Schools (JPS) and provides recommendations, based on ongoing discussions with the district and review of information from the benchmark assessments, on topics related to building a valid benchmark assessment system. We consider how item data can be used to improve benchmark tests over time and also cover what needs to be done to insure that results of a test are correctly interpreted, reported and used.

This report represents the second of two deliverables provided to the Jackson Public Schools (JPS) district on June 30, 2006, the first of which focused on assessment development. The distribution and assembly of items into a test is as important as writing good-quality items. The second deliverable, seen in the contents of this report, presents recommendations on the following specific topics:

I. Assessment item storage and distribution
II. Assessment administration
III. Assessment scoring, validation, and analysis
    a. Scoring
    b. Validation
    c. Analysis
IV. Assessment data use
    a. School and district evaluation and improvement
    b. Individual student performance
V. Documentation and communication of assessment findings
VI. Skills and competencies requirement of district staff

Assessment Item Storage and Distribution

During the assessment item development stage, the wording of the items should be checked and approved by the assessment design team, and the targeted content area and an estimated difficulty level should be assigned to each item. Each item should have a unique name and preferably be saved as an individual test item file. The targeted content area and the difficulty level of the item may be coded in the file name.

At the end of the assessment development stage, a summary sheet of all items developed should be prepared. The following lists some of the information typically included in the summary sheet:

- Grade level
- Subject area
- Content strand
- Name of test/s that include the item
- Type of item: multiple choice, open ended, etc.
- Actual item text
- Correct answer key for the item
- Dates of any pilot testing
- Dates it was used in district testing

A summary sheet prepared as above will help the assessment design team to distribute equivalent items into parallel test forms, use equivalent items for different purposes (for practice, for use in makeup exams, etc.), decide the best combination of test items for a specific assessment purpose, evaluate the content coverage of tests, etc. See Appendix A for an example of a summary sheet.

A test item summary sheet helps to ensure test security, enabling the assessment design team to create different but equivalent forms. This prevents students from cheating by obtaining a copy of a previous test. A test item file also helps the assessment design team to improve specific test quality by changing items that are not performing well, according to analyses of test results. In addition to constructing the test itself, a modified test item summary sheet, specific to a test, could also be given to teachers to make sure the teachers teach all the contents to be measured by the test ahead of time.

The distribution and assembly of items into a test is as important as writing good-quality items. Even the best test items can provide misleading or inaccurate information on student learning if the items are not assembled in the right way. There are three essential steps within the assessment assembly stage: reviewing test items, formatting the test, and preparing directions for the test (Worthen, Borg and White, 1993).

Reviewing Test Items

During the assessment assembly stage, the assessment team should: study the school or district assessment blueprint or guidelines that indicate the targeted content areas to be tested, examine the item summary sheet for the appropriate content coverage and item difficulty level, take into consideration what the student would have had an opportunity to learn before taking the test, think about the scheduled test time, and consider the implications for administration procedures.

Formatting the Test

A test consists of a set of test items organized in a logical way. It is important to assemble the test items in the appropriate manner. To assemble a test, one must first group together items with the same format. This will make the test instructions easier and students can answer the questions more efficiently. Next, the items should be presented from the easiest to the hardest. Presenting a few easy items that everyone can answer at the beginning of the test allows students a warm-up in taking the test and promotes confidence in answering later items. Many students become discouraged or give up when difficult items occur early in a test.

Another important point is that the test should be made as easy to read as possible. Items should have clear illustrations and enough spacing in between for students to distinguish one item from another. Drawings and figures (graphics) should be labeled clearly and put in the correct position in relation to the questions. Graphics should appear below the directions (if any) and above the question stem and response alternatives (Osterlind, 1998). In addition, predictable response patterns should be avoided so students focus on answering the question instead of figuring out the pattern of answers without knowing the contents. (For more on this point, see the item writing guidelines included in the first report.)

Preparing Directions for the Test

A good test should have clear and concise directions that all students can follow easily. It is often helpful to read the test directions aloud to the students, although directions should always be printed in the test booklet for students to refer to during the test. Test directions should explain what the students should do and how to do it. Additionally, a test should also include a specific set of directions for each type of item format, such as how answers will be recorded and how many points each item is worth.

The directions should also include information on the amount of available time for the whole test and for each section, so that students can pace themselves accordingly. The test direction should be reviewed by someone other than the test-writer to make sure everything is clear and easy to follow. Appendix B is a checklist for assembling of a good test (adapted from Worthen et al., p. 291).

In addition to the information delineated above, the directions should also describe the test purpose and its intended population, as well as how test scores should be interpreted and used. For example, any assumptions about student language ability, prior knowledge, format familiarity, appropriateness for special education students and any other groups for whom the test may not be appropriate, should be outlined. These points are addressed in more detail in later sections.

Extensive research has shown that students' opportunity to learn is directly related to performance on tests. It is important then to make sure that students not only have been given a chance to master the content to be tested, but also that they are familiar with the test procedures, test type, test format, and test duration. Whenever possible, practice tests should be offered to provide students with relevant testing experiences and to help ensure that later "official" tests will reflect students' true knowledge and ability with minimal error.

Assessment Administration

Assessment administration procedures and directions should describe the qualifications of administrators, administration procedures, and permissible variations in procedures, directions for administrators and test takers, and time limits. The following is a list of commonly accepted standards most relevant for test administration at the school level (AERA, APA, NCME, 1999):

- Test administrators have to be informed of the testing procedures and become familiar with the procedures before administering the test.

- Test administrators should follow carefully standardized procedures for administration as specified by the test developers.

- Test administrators should observe specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment.

- Test takers should be informed of the procedures for requesting and receiving accommodations, if these procedures have been established.

- Test takers should be instructed on how to respond to the questions. If there is equipment unfamiliar to the test-takers, they should be given clear instructions on the use of the equipment, as well as the opportunity to practice using the equipment.

- Schools should ensure a reasonably comfortable testing environment with as little distraction as possible.

- All personnel who have access to assessments should be aware of their responsibility in keep the test materials absolutely secured.

- Schools should also make efforts to ensure the security of the test and prevent any test takers from using fraudulent means to obtain higher scores. For example, in some state and district testing contexts a fixed number of test copies are given to each teacher and each class and the teacher is required to return all copies of the test, both those with student responses and those that are not used. No one is allowed to keep or make a copy of the test.

- If the standardized test administration procedures or scoring have to be modified to adapt to specific situations at school, the changes should be documented in a test report and possible effects of the modifications to test validity should be evaluated.

Assessment Scoring, Validation, and Analysis

Scoring

Test makers should develop a scoring procedure before the test is administered to students. If a test is made up of several parts dealing with different types of items or different content, a separate score on each part and a composite score on the whole test should be determined. This scoring can involve differential weighting of the score for each type of response. The typical item types are: a) selected response items such as true/false items, matching items, multiple choice items, and b) constructed response or open ended items such as completion items, short answer items and essay items (Kubiszyn & Borich, 2002). Each type of item requires a different approach to scoring. Scoring of multiple choice and open-ended questions is addressed here in detail, given that these predominate in JPS's current benchmark tests.

Multiple choice and other definitive answer items should be preferably scored by using a scanning machine to increase efficiency and reduce human errors. Care must be given in preparing and implementing the scoring procedure, and in documenting any variation. The test scoring group should monitor and report to schools and districts the frequency of scoring errors. Any systematic source of scoring errors should be corrected as soon as possible. In addition, the sources, rationale, empirical basis, as well as the limitations for computer-prepared interpretations of test response protocols should be discussed explicitly in a scoring manual.

For open-ended questions or performance assessments, raters need to be recruited and trained in order to score student responses correctly and reliably. Scoring should be based on how well students answer the question instead of, for example, hand writing (except on a test of penmanship). A detailed rubric or scoring guide should be provided for the raters/scorers. There are two kinds of scoring rubrics:

- Analytic rubrics distinguish different components of the answer and give credit for answering each component correctly.
- Holistic rubrics assess student work as a whole and give a general guideline on the overall quality of the answer.

These two types of rubrics assess student performance from different perspectives, and which type of rubric to choose depends on the nature of the test, cognitive level of students, and the complexity of desired responses. For example, younger students might not be able to attend to many details specified by an analytic rubric, so a holistic rubric might be a better choice. On the other hand, if the tests are going to be scored by teachers who might be inclined to implement their own scoring criteria, an analytic rubric would be advisable as it gives teachers stronger guidance and promotes uniformity in scoring.

Scoring sessions should be conducted where all the raters gather to discuss the scoring rubrics, score some of the test papers to identify key points in scoring the specific test, and reach agreements on difficult issues. After a fair amount of papers are scored, the consistency between the raters should be checked by calculating the percentage of exact agreement on the ratings as well as the closeness of the overall scores given by different raters on the same test. A more complicated inter-rater reliability coefficient can also be calculated to show the extent to which two or more raters agree in scoring the items. If the consistency is low between the raters, more training is needed to improve the situation until a fairly high consistency is reached. Or, if permitted, some raters with low consistency with other raters in general, should be removed from the scoring process.

A comprehensive set of scoring documents should be prepared before the actual scoring session, including documents like the following:

- Scoring rubrics
- Procedure for determining proficiency levels, criterion and/or norming procedures, and scaling procedures, if any
- Examples of anchor papers (i.e., papers exemplifying the lowest level of performance at each score point)
- Possible differential weighting of items
- Instruction and possible training on how to interpret and use the scores

Beyond the actual scoring process, if there is a material error detected in test scores the erroneous scores should be corrected and correct results distributed as soon as possible to all known recipients in order to avoid decision-making based on the erroneous scores. Confidentiality of the scores should be ensured when transmitting individually identified test scores to authorized groups or individuals, especially when transmitting by electronic media such as computer networks or facsimile. The actual test protocol and any written report should be kept in some form together with the test data about a person for future reference and quality improvement, and data should not be distributed without the context of the protocol.

Finally, as discussed in more detail later in this report, the meaning of scores should be clearly stated, and justification for score interpretations provided as part of the standard documents on assessments. Test score interpretations can become obsolete over time. A clear set of policy guidelines should be developed on how long the individual test scores should be retained, and how to make the scores available for longitudinal or other kinds of analysis.

Assessment Validation

Assessment development is an ongoing cycle of item development, revision, empirical testing, and validation. Modern conceptions of assessment validation require that evidence be collected to support the intended purposes and interpretations of an assessment. For example, a test designed to predict performance on another test should correlate highly with that test. (Thus JPS's assessments should correlate highly with state tests; see below for more information on predictive validity).

Validation evidence can include expert judgment, empirical studies, and statistical analyses. The reliability of an assessment should also be investigated, including internal and test-retest reliability and this reliability should be appropriate to the assessment's purposes. As noted in the first deliverable to JPS, other important validity considerations include fairness, alignment, and utility (Herman & Baker, 2005). Some of these topics were discussed in the initial deliverable to JPS, so in this report we provide more detailed information on other specific aspects of validity and reliability. To present this additional information, we use a traditional validity framework, which comprises: criterion-related validity, content validity, and construct validity (Allen and Yen, 1979; Brown, 1983; Worthen, Borg and White, 1993).

Criterion-related validity. For JPS, criterion-related validity is most important, since the ultimate purpose of the benchmark assessments is to predict students' later performance on the state assessment. This specific type of criterion-related validity is called predictive validity. Predictive validity indicates how well the new assessment predicts student performance on an established test. It is especially important in selection and placement decisions, such as college entrance exams. Predictive validity is typically measured by determining if student performance on the new test predicts their later performance on the criterion test, in this case, the Mississippi State test. Please refer to Wang, Niemi, and Wang (in press) for an example of conducting such a predictive validity analysis.

More generally, criterion-related validity examines how results of one test are related to some other external test or task that is supposed to measure the same content. In other words, how one can infer from a student's performance on one test, his or her performance on another external measure, which is called the criterion. In an educational setting, the criterion is usually an already existing test. For criterion validation, a representative sample of students are administered both the criterion test and the current test. Then a correlation coefficient is computed between the results of the two tests for the same students to summarize the extent to which the two tests are correlated. This coefficient is called a validity coefficient.

The external criterion test can be taken at the same time or later than the new test. If the two tests are taken at the same time, the concurrent validity of the new test is examined. Concurrent validity is usually used to check if a new test, such as a teacher-made test, is a good alternative to a more expensive measure. As described above, when the criterion test is taken after the new test, it is used to examine the predictive validity of the new test.

Content validity. Content validity examines how well the items from a test represent the entire content domain to be measured. Content validation involves directly comparing the items with the domain they are expected to assess. If the items on a test accurately and comprehensively represent the objectives to be measured, the test is considered to have content validity. A test with good content validity should comprehensively represent both the subject-matter topics (such as reading or mathematics, domains that are tested) and the cognitive processes that students are expected to apply to the topics.

There are four basic steps to evaluate content validity. The first step is to specify the domain of the assessment as clearly as possible and divide it into subcategories with importance levels attached to each of them. (This process was discussed in the first deliverable.[1]) Then the number of test items for each subcategory should be evaluated with respect to the total test length and importance of the subcategories, and also to make sure that all content areas and instructional objectives are represented. To accomplish this goal, we recommended specific alignment procedures in Deliverable 1.

[1] For further information please see CSE Technical Report 723.

Content validation reviews should also address whether test items introduce error not related to the content being tested, such as a clue in the question that enables students to answer the question without having the content knowledge, or a poorly worded item that enables students to answer incorrectly even if they have the content knowledge (Gronlund & Linn, 1990).

Construct validity. Construct validity refers to the extent to which a test measures the construct it is supposed to measure. A construct is an unobserved, postulated human attribute or concept that cannot be directly measured (Worthen, et. al., 1993), for example, "understanding how addition and subtraction are related". Other examples of constructs include: children's readiness to learn how to read, students' anxiety about learning physics, or understanding of algebra. Construct validation requires collection of empirical data from different sources to analyze if a test measures the constructs it proposes to. The evidence can include, but is not limited to, the aforementioned content validity and criterion-related validity.

In general, construct validation involves the following steps:

- Define the construct domain to be tested.
- Describe how the test represents the important aspects of the domain.
- Examine how students' performances on the test are related to their performance on other measures of the same construct (e.g., relevant items on the state test).
- Examine if students' scores on the test have relatively low correlations with measures that are supposed to be uncorrelated with the construct being assessed.

The first two of these steps were covered in Deliverable 1. The third step could be accomplished by conducting the correlational studies described earlier. The fourth step might be accomplished by examining correlations with other subject area tests, e.g., the correlation between a math and a reading test.

Reliability

Reliability addresses the extent to which a test consistently measures what it is supposed to measure and informs on how well the estimated test score reflects a student's "true score" on a test. Test scores are reliable if they reflect the true score of each student. A true score is the accurate assessment of a student's ability or performance on what is being tested. However, there are always other factors (errors) that can contribute to a student's observed test score, such as the influence of a student's health or the testing environment at the time when the test is taken. For example, if a student takes a test twice, most probably he will not have exactly the same score both times due to random or non-random factors that influence his test performance. These errors cause him to score higher or lower than his true score and are called measurement errors. Therefore, an observed test score actually includes two parts, the true score and measurement error.

There are two kinds of measurement errors: systematic error and the random error. Systematic error is usually uniform, with the magnitude and the direction of the error the same for the test-takers. For example, if there is a typo in the test which distorted the whole meaning of an item, all student performance will be affected in the same way due to that. That item's score would not be a reflection of the knowledge or skills supposedly tested by that item. Systematic errors reduce the validity and the meaning of the test scores. Careful design of test items and control of test administration can help avoid systematic error. Random error refers to unsystematic errors, which differ from one test-taker to another and make the test results inconsistent and unreliable. The key point in achieving high reliability of a test is to reduce random errors by controlling the extraneous factors that cause them. In other words, the goal of reliability investigations is to estimate the consistency of scores produced by a test and the extent to which the test score is free from random error.

The various methods used to estimate the reliability of a test include test-retest reliability, parallel-form reliability, Cronbach's alpha method, split-half method, and the Ge-Richardson method. Test-retest and parallel-form reliability both involve testing the same students twice, and are not very feasible in many situations. Examining the reliability using a single test administration is more popular. Of the three measures of internal consistency—Cronbach's alpha method, the split-half method, and the Kuder-Richardson method—Cronbach's alpha method is the most commonly used and is recommended to JPS. The Kuder-Richardson method may also be of use to JPS. All three of the internal consistency measures are described in more detail below.

Split-half method. The split-half method measures the reliability of a test by correlating one half of the test with the other half. The key point here is to split the test into two equal halves that are similar in content and difficulty. For a test where items are ordered by difficulty, the odd-even split is generally used to separate the test into two parts. If the items are not ordered by difficulty, the test can be divided into the first and last halves or the items can be assigned randomly into two halves. No matter what method is used, careful examination should be conducted of the two halves of the test to see if they are in fact parallel.

A general point to keep in mind is, if all items on a test measure the same construct, a longer test is more reliable than a shorter test. Because the weight given to any one single item is lower in the longer test, differences in responses on one item will not have as much impact on total score than a shorter test. In the split-half method, the correlation coefficient is based on the two halves of half the length of the test. To make this correlation coefficient comparable to the coefficient obtained from the test-retest or parallel-form method, the correlation coefficient must be transformed using the following formula:

$$R=2r/(1+r)$$

where R represents the corrected reliability coefficient and r is the original correlation coefficient between the two halves of the test.

The split-half method has obvious advantages over the test-retest or parallel-form method in that it does not have the problem of memory or practice effects, nor is it subject to differential administration or scoring. However, it cannot be used in speeded tests (where students are given limited time) like the other two methods. Speeded tests usually contain many items that are easy and students usually answer the early items right and omit the later items, inflating the reliability coefficient.

Kuder-Richardson method. In the split-half method, the resulting reliability coefficient depends on what kind of splitting is performed. The reliability coefficient calculated from odd-even splitting will be different from that calculated from first-and-last splitting. To solve this problem, Kuder and Richardson (1937) proposed to calculate the average reliability of all possible splitting methods for a test. In this way the Kuder-Richardson method avoids the actual split of a test into two halves, while maintaining the advantages of split-half methods: only a single test administration is needed, and the reliability coefficient is not influenced by memory or practice effects.

There are several drawbacks to the Kuder-Richardson method, too. Like the split-half method, it cannot be used in speeded tests. In addition, it can only be used for tests with items that are dichotomously scored (two possible answers or scores), and all the items in the test have to measure the same construct. For example, the Kuder-Richardson method cannot be used on a math test that tests different dimensions of math, such as a test measuring student ability, multiplication and division of fractions, and problem solving, all together. The Kuder-Richardson reliability coefficient will underestimate the reliability of a test with multiple dimensions.

Cronbach's alpha method. Cronbach's alpha method estimates the internal consistency of a single test, especially when the items are not all dichotomously scored. It is particularly popular for estimating the reliability of tests with essay and short-answer questions. Similar to the Kuder-Richardson method, it estimates the average correlation of all possible ways of splitting a test. The computation is complex but can be performed using computer software, such as SPSS or SAS.

Like the Kuder-Richardson method and split-half method, Cronbach's alpha method does not have the problem of memory or practice effects. Although it cannot be used in speeded tests, it can be applied to more types of tests than any other methods that are available.

Parallel-form reliability. Parallel-form reliability refers to the extent to which students' scores on two or more "equivalent" forms of the same test agree with one another. If the test has high consistency, the correlation between the scores of the same students on different forms of the test should be high. Compared to the test-retest method, the parallel-form method avoids possible memory effects. However, the key to the parallel-form reliability method is that different forms of the test are equivalent in content, form and difficulty, but not so similar as to have essentially the same items (e.g., 5+6 versus 6+5). Memory of identical or similar items from the first test can affect student performance on a second test, which is a very difficult effect to avoid. For this reason, parallel-form reliability is typically only tested by large test publishers with sufficient resources.

If parallel-form reliability is tested, the parallel forms of the same test should be administered between short time intervals, to minimize learning between the tests. There are different opinions regarding optimal time intervals for this method. Some specialists suggest that no interval be given between the tests to reduce errors due to day-to-day variations within individuals. While others argue that the test results themselves will be unreliable due to factors such as fatigue, memory and practice effect, and sufficient time should be given before students take the second form of test. Like the test-retest method, the parallel-form method can be used for speeded tests. Since parallel reliability coefficients are usually lower than test-retest coefficients, special care must be given when interpreting the coefficients. Another method for creating parallel forms is to use item-response theory (IRT) to equate different forms; this requires that some but not all of the items be repeated on both forms of the test. Conducting IRT analyses requires a high level of technical expertise, and JPS may want to consult with IRT experts if it chooses this approach.

Test-retest reliability. Test-retest reliability is estimated by the correlation between student scores from the same test that is administered twice. If the test is not influenced much by random errors, individual students should get very similar scores on the two tests and the correlation between the two scores should be high. The advantage of this method is obvious. There is no random error related to difference in test items or the format of the tests since exactly the same test is taken twice. In addition, there is no need to construct an alternative form of the test as is the case in some other methods and the reliability coefficient is easy to calculate. However, there are several problems with this method that we should note.

First, by using this method, students have to take the same test on two different occasions, which might not be feasible in many situations. The second problem is memory effects, described above. If the same test is administered to the students shortly after they take it the first time, many students will be able to recall their answers to the items in the first test and will just fill in the answers without active thinking. This will inflate the estimated test reliability. The third problem arises when the same test is administered to the same students after a long time interval. Although memory might not play an important role after an extended interval, students might acquire new knowledge or skills. In this case, students will perform better when they take the test the second time, and the estimated test reliability will appear to be lower than it actually is. Due to the drawbacks of both short-interval and long-interval testing, it is important to find an appropriate time interval between the two tests based on the content and population to be assessed. Test-retest reliability is good for speeded tests, when there is only limited time for testing, or when there are no parallel forms.

Fairness

There are many issues relating to the fairness of testing. Fairness can be interpreted as equitable treatment in the testing process, equality in outcomes of testing, and opportunity to learn (American Educational Research Association et al., 1999). Providing all examinees with a comparable chance to demonstrate their knowledge and skills on what the test measures is one of the aspects. Fairness considerations include such factors as equal familiarity with the test format for all students, proper testing conditions with standardized tests and administration conditions, and accurate and comprehensive reporting of individual or group test results.

The fairness of a test can be examined by comparing the new test with other well-established tests. A fair assessment should be an accurate reflection of student mastery of the contents, and should not be unduly influenced by students' background characteristics. For example, Wang, Niemi and Wang (in press) examined whether a writing performance assessment is a fair measure of student proficiency by looking at both variance partition and the effects of student and school variables on student achievement on both the performance assessment and the Stanford Achievement Test (SAT9) reading and mathematics scores. They found that the performance assessment might be a more egalitarian test than the SAT9 tests, because student background variables, such as student ethnicity and family SES, had less explanatory power for the performance assessment than for the SAT9 tests; that is, on the performance assessment a greater proportion of variance in student scores was accounted for by student knowledge rather than background characteristics.

Language fairness is also an important testing consideration. Testing of individuals with insufficient skills in the language used on a test can lead to scores that do not accurately reflect the construct measured by the test; this could happen if students with high math knowledge were unable to read the instructions or word problems on a math test, for example. It is thus important to take language background into consideration in the development, selection, and administration of the test, as well as in the interpretation of the test results. Specifically:

- The threats to the reliability and validity of test score inferences due to language differences should be addressed in designing the test using professional judgment together with empirical studies.

- If there is credible evidence that the test score is biased across subgroups of linguistically diverse test takers, separate validity evidence should be collected for each subgroup as well as for the whole population of test takers.

- If a test taker is proficient in two or more languages in which the test is available, the test taker should be administered the test in the language in which he or she is most proficient (unless language proficiency itself is part of the test objective).

- The test manual should describe in detail the linguistic modifications recommended for the test and the rationale for the modifications.

- Appropriate test use and interpretation should be provided by the test developers and publishers if a test is to be used on linguistically diverse test takers.

- If a test is translated from one language to another, the methods to ensure the adequacy of the translation should be described and evidence for score reliability and validity of the translated version should be presented.

- If there are multiple language versions of a test, the evidence of test comparability should be presented.

- If an interpreter is necessary in testing, the interpreter should be fluent in both languages, be proficient in translating, and should be familiar with the assessment process.

(Adapted from standards listed in AERA et al., 1999, p. 97.)

The other two important aspects of fairness in testing are cultural fairness and gender fairness. Cultural fairness has to do with whether students from different ethnic groups have the same performance on the test, controlling for their true ability level and other confounding factors. For example, if a reading test contains materials that are related to the mainstream culture of European-Americans, African-American and Asian-American students might not perform as well as European-American students on the test, even if they have the same reading ability.

A test may also contain gender bias. Male students have been found to perform better in multiple choice tests and female students have been found to achieve higher scores on essay and oral exams. Cronbach (1976) also found that male students scored higher on nonverbal items and female students did better on verbal items. Nevertheless, these differences are not substantial and might be remedied by training. In addition to cultural and gender factors, fairness can also be related to student's socio-economic status (SES), since students with higher SES tend to have more home resources than students with lower SES.

In considering all of these fairness-related issues, Thorndike (1971) pointed out that it is not the test itself, but the use of the test that can be fair or unfair. Fairness is not a major concern when the purpose of a test is to identify strengths and weaknesses in order to improve student achievement. On the other hand, fairness should be a concern if the test is used for academic decisions about individual students, such as ability group placement or college admission, or if it is used to evaluate schools and teachers.

Assessment Analysis

Item analysis is the process of examining test items individually. (Several examples of item analysis were presented in Deliverable 1.) It provides a powerful means to judge the quality of test items and to improve the validity and reliability of the test as a whole. This section focuses on the numerical analysis of student responses, including computing item difficulty level and item discrimination indices, evaluating the effectiveness of distractors, factor analysis, item analysis using reliability and validity indices, and item analysis with item-characteristic curves (ICCs; Allen and Yen, 1979; Baker, 2001; Osterlind, 1998). Of these, it is most important and informative to evaluate the individual item difficulty level, to use factor analysis and reliability indices to make sure all items are necessary, and to use factor analysis to verify the number of constructs the test is supposed to measure.

Item difficulty. Item difficulty level is represented in the percentage of students who answer an item correctly. This is also called the item's p-value. Items on norm-referenced tests should have p-values of .20 to .80, indicating 20% to 80% of students answered the item correctly. If an item is answered correctly by almost none of the students or nearly all of them, it does not provide much information about the knowledge or skills assessed by this item.

In terms of multiple-choice items, p-values can also provide information about an item's general performance and pinpoint mistakes in writing, such as a wording mistake that makes an item confusing to the test-taker, misleading distractors, or other problems. For a multiple-choice item, a p-value can be computed for each of the possible responses to show the percentage of students who choose that answer. For example, suppose 30%, 0%, 40% and 30% of the students choose answers A, B, C, and D respectively for an item, with A as the correct answer. The p-values for answers A, B, C and D are .3, 0, .4 and .2, respectively. Such widely scattered responses may indicate that there is a problem with the item. The correct response, A, is chosen by fewer students than response D, which is supposed to be a distractor. Response B is obviously an unsuccessful distractor since no students chose it. Also of note is that the p-values of all the responses add up to .9, which indicates that 10% of the students did not answer the question. This further suggests that this item may be confusing to the students, and it should be reviewed and possibly revised or removed from the test.

Item discrimination. The item discrimination index identifies how well an item distinguishes high-achieving students from low-achieving students on a test. If an item has a much higher p-value for students who have high overall scores than those who have low scores on the test, this item is considered to have high discrimination power. On the other hand, if an item has a similar p-value for students who have high scores and those who have low scores on the test, this item is considered to have low discrimination power. A discrimination index is scaled from –1.0 to +1.0. A positive discrimination index indicates that if a student gets this item right, he is very likely to get a higher total score on the test. A negative discrimination index suggests that this item does not measure the same construct as the test in general, and should be revised or taken out of the test.

In addition to checking p-values on students grouped by high and low achievement levels, one can also divide students into more categories to examine and compare the p-values of these categories. For example, if there are a large number of students taking the test, they can be divided into five sections, each representing 20% of students in the score distribution. A graph can be constructed to check if the general trend of percent of students answering an item correctly increases as student scores on the total test increase.

Effectiveness of distractors. In a multiple-choice item, distractors should be plausible but incorrect answers to test if students have mastered the concept being assessed. Distractors not only provide information about students' misunderstanding of concepts, but also suggest why these misunderstandings occur. The evaluation of distractors involves the calculation of the p-value for each of the answers for different achievement groups. If there are sufficient students taking the test, each answer option should be chosen by at least a few students. An answer option that is not chosen by any student is not a good distractor since it provides no information about student knowledge or skills on the tested concept. Students in the low achieving group should choose the distractors more frequently than students in the high achieving group. If a distractor does not distinguish high achieving students from low achieving students, it impairs the discrimination power of the item.

Item analysis using factor analysis. Factor analysis studies the inter-correlation between the items in a test and determines the theoretical constructs that might be represented by the set of items in a test. There are two kinds of factor analysis: exploratory factor analysis and confirmatory factor analysis. With exploratory factor analysis, researchers examine the output from a principal component analysis and then decide on the number of factors underlying the set of measures. By allowing all items (measures) to load on all factors, exploratory factor analysis examines the factor structure and the internal reliability of the factors. In confirmatory factor analysis, specific factor structure is defined by a prior theory, where the researcher indicates which items load on which factor.

If a test is supposed to measure different constructs, exploratory factor analysis can be used to determine what construct each item measures and how well it measures the construct. On the other hand, confirmatory factor analysis should be applied if a test is expected to measure only one construct. Confirmatory factor analysis determines how well the item measures the construct through its loading on the factor. A factor loading is similar to a correlation coefficient between the item and the factor, which is not observed, but measured. The factor loading of an item on a construct indicates how well that item measures the construct.

Item analysis using reliability and validity indices. Often, a test developer needs to choose a certain number of best items from a larger number of available test items to compile a test. In order to decrease the time needed for testing, it often makes sense to choose the least amount of items that have the same internal consistency as larger item sets. Under these situations the item reliability coefficient can be used to select items. For example, if a test developer wants to choose 10 items from a pool of 15 items to make a test, the developer first runs a reliability analysis on all 15 items. In addition to a general reliability coefficient, an item-reliability index is also obtained, which represents the correlation between the item score and the total test score controlling for the standard deviation of the item score. Items with a higher item-reliability index should be selected. After the developer selects the 10 items based on the item-reliability index, she/he conducts another reliability analysis with only these 10 items and compares the results with the original reliability coefficient with 15 items.

In an alternative approach, if the test developer wants to choose the least amount of items with similar internal consistency, she/he may initially eliminate the item with the lowest item-reliability index. Next the developer will determine the reliability of the remaining 14 items to check if removing the item improves the reliability of the test. Then she/he should remove the item with the lowest item-reliability index based on the 14 item test. This process should be repeated until the reliability does not improve, or decreases, when additional items are removed.

Item validity should also be considered in selecting items. In validating of the test, an item validity index can be obtained for each item based on the correlation between the item score and the criterion score. The validity index and reliability analysis should both be evaluated to make a test with the highest validity and reliability. However, a test with the highest possible reliability might not have the highest possible validity. A final decision on item selection has to be made with practical consideration of the nature and the purpose of the test.

Item analysis with Item-Characteristic Curves (ICCs). An item-characteristic curve (ICC) is a graphical presentation of the relationship between the difficulty of an item and the extent to which the item measures the underlying trait that is measured by the test (Allen & Yen, 1979). Modern ICC also includes the probability of guessing a correct response for examinees with very low abilities (Baker, 2001). The advantage of the ICC is its ability to graphically present several characteristics of an item simultaneously in an easy-to-understand fashion. The following figure shows a typical ICC. The Y axis of the plot represents the probability of a correct response and the X axis represents the proficiency of students on the latent trait, which ranges from –3.0 to +3.0. The position of the curve on the X axis informs on the difficulty of the item and the steepness of the curve shows how well the item discriminates between students with high proficiency and students with low proficiency. The intercept of the ICC and the Y axis indicates the probability of guessing the correct answer for a student with minimum proficiency.

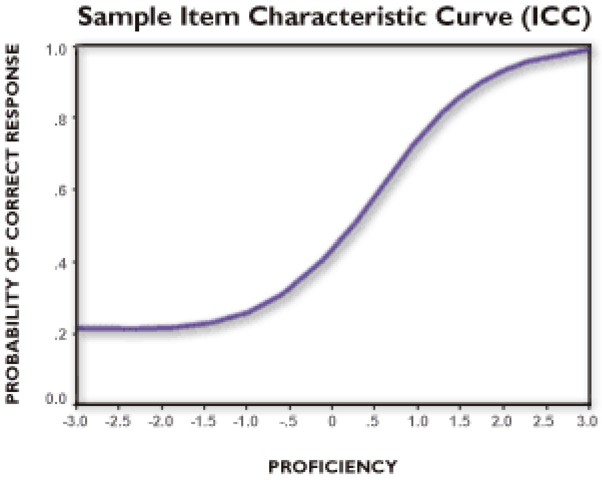


Figure 1.

On an ICC graph with all items, if the ICC of Item A is to the right of the ICC of Item B, Item A is more difficult than Item B. If the ICC of Item A is steeper than the ICC of Item B, Item A distinguishes students with high proficiency from students with low proficiency better than Item B. ICCs are usually estimated using item response theory, which examines the latent traits, such as examinee characteristics or hypothetical constructs, which cause students with a certain level of proficiency to perform similarly on a test or an item. Since IRT uses complex procedures to evaluate difficult concepts, it is not discussed in detail in this report, but it nevertheless is a recommended methodology.

ICCs can be used to detect item bias, i. e., whether the probability of success on an item is not the same for student subgroups at the same proficiency level. Examples of subgroups to consider are gender, ethnicity, socio-economic status, and language status. If students from one subgroup consistently perform worse on an item than students from other subgroups who are at the same proficiency level, it is said that this item is biased toward this subgroup. For example, a reading comprehension item on farming may work differently for students from rural areas than students from urban areas. If we plot the ICC of such an item separately on the two subgroups, the resulting two ICCs may look different. The item may have better discrimination power in the rural subgroup than the urban subgroups. The ICC plots reveal not only that the item performs differently for the two subgroups, but also the nature of the differential performance.

Based on ICC analysis, the test developer can revise the item to make it easier or more difficult, make it more discriminating among different proficiency levels, or try to reduce the effects of guessing.

Assessment Data Use

School and District Evaluation and Improvement

There are usually many testing activities going on in a school, with different tests serving different purposes. Teachers use tests to determine grades, or sometimes to estimate how well the students have mastered what was taught during class in order to adjust instruction and help students with learning needs. Districts and schools administer tests for structural, instructional management, and student placement decisions, as well as for evaluation and accountability purposes. State tests typically evaluate whether schools are meeting the standard. In addition, there are national level tests that are used to evaluate progress in broad areas such as mathematics achievement, or to evaluate entitlement programs in education.

Without good organization and planning, these multiple tests may be redundant as well as inefficient in assessing the most important educational objectives. Districts and schools should develop integrated testing programs consisting of a reasonable combination of classroom level and larger-scale tests. Plans for these programs should address questions such as how different assessments complement or supplement each other, how information from different assessments can be combined, and how information from all the assessments can be used effectively. In the ideal case, the assessment system will be able to inform on student learning, instruction and curricula, and district and school administration, as well as providing information to identify and solve educational problems (Lyman, 1998; Traxler, Jacobs, Selover and Townsend, 1953; Worthen, Borg and White, 1993).

The first step for a good testing program is to identify the purposes of the tests (a process described more fully in Deliverable 1. Each district has its own goals, organizations, and curricula. It is important for the district decision-makers to carefully think through the real purposes of the testing based on the unique features of the district. For example, if the pressing needs in a district are to improve the achievement of low-ability students, it should include minimum competency testing in order to identify these students. It is also important to select norm-referenced or criterion-referenced measurement. If the district intends to establish the relative performance of its students compared with the national average, it should choose norm-referenced tests. On the other hand, if the district wants to know if all students have mastered a set of well-defined standards, criterion-referenced tests are preferred. While it is essential to focus on the most important goals of the district, care should be taken not to neglect other goals. For example, if a district testing program over-emphasizes basic skills, teachers would correspondingly spend more time on teaching these basic skills and neglect other skills, including conceptual understanding and complex problem solving. Additionally, a good testing program should inform teachers and administrators of the district goals and whether these goals are achieved based on the test results.

For a testing program to be successful, it should involve participation from not only testing specialists, but also by teachers, educators, and possibly even parents and students. Including all stakeholders in developing or reviewing the testing program may increase its sensitivity to classroom and schools' instructional objectives and its acceptance across constituencies.

Using Assessment Data to Evaluate Individual Student Performance

There are a variety of ways to use assessment data to evaluate individual student performance, e.g., comparing students to an absolute standard or to each other. For JPS, considering its purpose in using benchmark assessments, the focus should be on comparing student scores to scores on the state test. The following paragraphs summarize the most popular systems used to judge student performance against a standard.

Comparison with an Absolute Standard

The most common method in this category is the percentage correct score, where each student's score on the test is compared with the maximum possible score. It is an easy method to adopt and the results are straightforward. However, the percentage correct score is test-sensitive; students will get different scores from different tests based on the difficulty levels of the tests. This score alone will not indicate how well the students have mastered the content area in general or compared with others.

Comparison with Other Students

When the purpose is to compare students to each other, the difficulty of the test becomes irrelevant because it is based on the relative proficiency of one student versus all other comparable students. When the test is more difficult, all students will have lower achievement scores than they would on an easier test, while students with better content knowledge or skills will obtain higher scores than students who are less knowledgeable or skillful.

A standardized score can be calculated to compare the performance of students with one another; the most basic standardized score is the z-score. The z-score suggests how far a student's score deviates from the general mean in terms of standard deviations. The z-scores follow the z-distribution, where about 68% of the scores fall within the range of – 1 to + 1 standard deviation below the mean, 95% of the scores fall within the range of – 2 to + 2, and over 99% of the scores fall between – 3 and + 3. Since it's symmetrical, around half of all the z-scores are negative, which makes it difficult to interpret. In addition, it is difficult for people not familiar with statistical concepts to make sense of a score from – 3 to + 3.

While derived from the z-score, the t-score is becoming one of the most popular standard scores. It has a mean of 50 and a standard deviation of 10. The t-score is easier to interpret since it does not have negative values. There is also the normal curve equivalent score, which is a normalized standard score with a mean of 50 and a standard deviation of 21.

Ranking is another technique used to compare students. The most popular rank score is percentile rank, which specifies a student's position relative to a group of students. The obvious advantage of the percentile rank score is that it is easy to interpret. However, the achievement difference represented by a 1 percentile difference is not the same across the whole scale, which can easily cause misunderstanding. Another variation of percentile rank is decile ranking, in which the frequency distribution of the scores is divided evenly into 10 groups. The first decile is the same as the tenth percentile, and the second decile equals the twentieth percentile, and so forth.

Documentation and Communication of Assessment Findings

While the Jackson Public School district's primary stated goal for using benchmark assessments is to predict scores on state standardized tests, the benchmark scores may also used to take action and make instructional change. Besides sharing student and school data with schools, district wide results can also be disseminated. District-wide results can help to identify areas in need of improvement; school and teacher-level results can reveal where changes in instruction are needed.

Assuming that JPS shares benchmark assessment results throughout the district, the following are recommendations regarding distribution of findings:

Timeliness

Test results should be distributed as soon as possible so there is time to implement recommended changes. This is particularly important if some content needs to be taught. Taking into account practical constraints, JPS should provide schools and teachers with as much time as possible to study assessment results and re-address the content if necessary. Ideally, benchmark results will be released for formative purposes on a set schedule throughout the year with a summative report prepared once the state test scores are available.

Tailoring Reports to the Audience

District administrators, schools, and teachers have different needs for assessment information. District administrators, for example, may be most interested in district-wide and school-by-school results with a focus on the relationship of benchmark tests to state assessments. As a result, they may prefer a smaller number of comprehensive reports. Teachers, on the other hand, may be more interested in how test results could be used to guide their practice with individual students and their classrooms as a whole. Thus, teachers would be better served by reports including individual student data that is tied to their curricular schedule, as discussed above.

Presentation of Data

As discussed later in this report, administrators and teachers should receive professional development support to understand the content and appropriate uses of assessment information. However, even with this training it cannot be assumed that merely putting benchmark assessment scores into a teacher's or administrator's hands is adequate to assure understanding and use of the data; additional training on subject area knowledge or pedagogical strategies may be necessary. Report results should be understandable even to those with minimal assessment or psychometric training. These reports should not only contain adequate explanatory information about the meaning of scores (e.g., what a scale score represents), but also highlight key findings and relative strengths and weaknesses across the curriculum. Results could also potentially be broken down into various subgroup categories reflecting the diversity of students in the district, although teachers may not be able to tailor instruction to accommodate every subgroup.

Support/Structure for Data Presentation

Again, even with some professional development and guidance in interpretation of scores, it cannot be assumed that teachers, or even administrators, will know how to respond to assessment findings. Some districts have taken the approach of providing concrete suggestions for instruction or other student support along with the assessment findings. Depending on available resources, this approach is not always feasible. As an alternative, JPS could direct teachers to materials already available in the district, such as curriculum guidelines and similar resources.

Additionally, many districts have assembled teams specifically to review assessment findings and plan responses to assessment results. These teams could be at the grade, school or district level, and meet on a regular basis, such as monthly or quarterly, throughout the school year. The teams can include teachers, content experts, and even administrators. Such regular opportunities for discussion and planning have proven very useful for devising next steps based on assessment results. Preferably, these meetings can be built into the existing professional development and meeting schedules of the district, rather than added on as an additional requirement or responsibility.

Required Skills and Competencies

The recommendations in this report imply additional competencies that go beyond those described in CRESST's first report to JPS. District and school staff may already possess many of these additional competencies, which include:

- Expertise in the development of assessment items and tests, including the design and processing of scannable documents (for JPS central office/administrative staff).

- Understanding of test administration procedures (for teachers and JPS or school personnel overseeing test administration).

- Knowledge of item scoring procedures including, if appropriate, use of rubrics (for test designers, scoring trainers, and any teachers expected to score student responses).

- Basic understanding of properties of tests and their measurement. Specifically, this includes the skills needed to conduct statistical analyses of test quality such as inter-item correlations, reliability, and item analysis (for JPS testing/assessment staff).

- Ability to conduct statistical studies of validity, including predictive validity vis-à-vis state tests (for JPS testing/assessment office staff).

- Skill in the design and creation of user-friendly assessment reports (for JPS testing/assessment office staff).

- Basic understanding of the meaning and interpretation of assessment findings (for teachers, and administrators at all levels).

- For each content area: knowledge of curricular/instructional implications of assessment findings (i.e., what specific instructional changes could be made in response to certain assessment outcomes). Although this knowledge applies to teachers, administrators, and the district central office, teachers should, at minimum, know where and how, in the district, to access the resources and information to help them make instructional decisions based on assessment results (for teachers and professional development staff).

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.

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

Baker, F. (2001). *The basics of item response theory*. College Park, MD: Eric Clearinghouse on Assessment and Evaluation.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: CBS College Publishing.

Cronbach, L. J. (1976). Equity in selection - where psychometrics and political philosophy meet. *Journal of Educational Measurement, 13*(1976), 31-41.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Herman, J. L. & Baker, E. L. (2005). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership, 63*(3), 48-55.

Kubiszyn, T., & Borich, G. (2002). *Educational testing and measurement: Classroom application and management* (7th ed.). New York: John Wiley & Sons Inc.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrica, 2*, 151-160.

Lyman, H. B. (1998). *Test scores and what they mean* (6th ed.). Boston, MA: Allyn and Bacon.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and other formats* (2nd ed.). New York: Kluwer Academic Publishers.

Thorndike, R. L. (1971). *Educational measurement*. Washington, DC: American Council on Education.

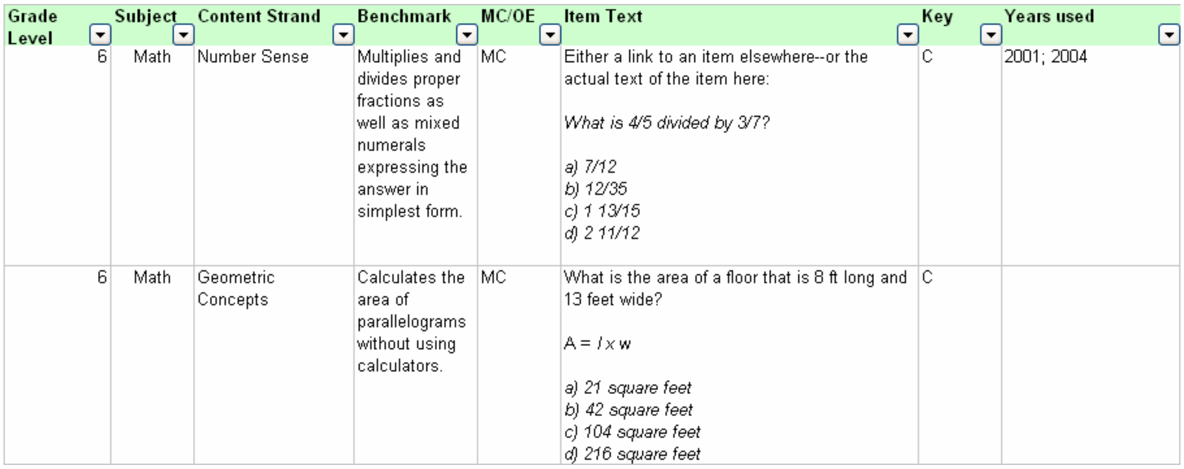
Traxler, A. E., Jacobs, R., Selover, M., & Townsend, A. (1953). *Introduction to testing and the use of test results*. New York: Harper & Brothers.

Wang, J., Niemi, D., & Wang, H. (in press). Predictive validity and fairness of an English language arts performance assessment. *Educational Assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools*. New York: Longman Publishing Group.

Appendix A

Sample Assessment Summary Sheet



Grade Level	Subject	Content Strand	Benchmark	MC/OE	Item Text	Key	Years used
6	Math	Number Sense	Multiplies and divides proper fractions as well as mixed numerals expressing the answer in simplest form.	MC	Either a link to an item elsewhere--or the actual text of the item here: *What is 4/5 divided by 3/7?* *a) 7/12* *b) 12/35* *c) 1 13/15* *d) 2 11/12*	C	2001; 2004
6	Math	Geometric Concepts	Calculates the area of parallelograms without using calculators.	MC	What is the area of a floor that is 8 ft long and 13 feet wide? A = *l x* w *a) 21 square feet* *b) 42 square feet* *c) 104 square feet* *d) 216 square feet*	C	

Appendix B

Checklist to Use in Assembling Items into a Useful Test

1. Are items clear, concise, and free from jargon or unnecessarily difficult vocabulary?
2. Do items avoid sexual or racial bias?
3. Has someone besides the item writers reviewed the items?
4. Has the answer key been checked?
5. Are items grouped according to type or format?
6. Are items arranged from easiest to most difficult?
7. Are items appropriately spaced?
8. Are item stems, options, and support material (e.g., diagrams) appropriately arranged?
9. Is the answer sequence random?
10. Are items representative of the material taught?
11. Can the test be completed in a reasonable time?
12. Have directions been checked for clarity?
13. Has the test been rigorously proofread for errors?
14. Is a space provided for students to write their names?