White Paper

Composition of a Comprehensive Assessment System

by
John Richard Bergan, Ph.D.
Christine G. Burnham, Ph.D.
John Robert Bergan, Ph.D.
Sarah M. Callahan, Ph.D.
and Jason K. Feld, Ph.D.

Assessment Technology Incorporated

Assessment Technology, Incorporated
6700 E. Speedway Boulevard
Tucson, Arizona 85710
Phone: 520.323.9033 • Fax: 520.323.9139

Assessment Technology, Inc., Publishers
Tucson, Arizona, U.S.A.
Printed in the United States of America.
V6-013013

Composition of a Comprehensive Assessment System

By John Richard Bergan, Ph.D., Christine G. Burnham, Ph.D.,
John Robert Bergan, Ph.D., Sarah M. Callahan, Ph.D., and Jason K. Feld, Ph.D.
Assessment Technology, Incorporated

Table of Contents

I. Introduction

The standards-based education movement has been accompanied by increased use of assessment information to guide educational decisions. The increased use of assessment calls for the development of comprehensive assessment systems capable of providing the data required to inform the many types of decisions confronting educators in today's schools.

Different types of decisions require different types of assessment. A comprehensive assessment system designed to meet assessment needs in standards-based education is composed of many types of assessment. Each type serves a different central purpose, and each type may require variations in test characteristics and assessment procedures in order to serve its central purpose effectively. Although each type of assessment has a unique central purpose, there is a considerable amount of overlap in subordinate purposes that the various types of assessment may serve. In some instances the overlap may lead to increases in assessment efficiency in that a given type of assessment may fulfill more than one assessment need. In other instances, the overlap will add additional information enhancing the effectiveness of educational decisions.

Comprehensive assessment systems hold the promise of providing beneficial changes in the ways in which assessment is used in education. Perhaps the most significant change is the increase that a comprehensive system can provide in the amount of information available to improve education. In a comprehensive system, results from a variety of assessments can be placed on a common scale. The use of a common scale makes is possible to directly compare results from different assessments. The availability of multiple sources of comparable information offers a more complete picture of academic accomplishments than would otherwise be possible to achieve.

A comprehensive assessment system provides the additional advantage of increasing assessment efficiency. In ATI's comprehensive system, a technology platform common across assessment types is developed to facilitate implementation of the system. The platform includes artificial intelligence routines that guide test construction, test review, test publication, test security, test scheduling, test administration, test scoring, psychometric analyses, reporting, and the provision of actionable decision options.

A third advantage of a comprehensive system involves system monitoring tools that make it possible to guide system implementation. These tools include information regarding when tests have been constructed, reviewed, revised, and delivered. In addition, they include data on test administration such as the students scheduled to take the test, the number who have actually done so, and when testing has been completed. Monitoring information also provides data on the extent to which different types of assessment are used. This information can be used to guide item development.

This paper outlines the composition of a comprehensive assessment system designed to inform the varying types of educational decisions that must be addressed in schools today. The paper details the major types of assessments included in a comprehensive system, the purposes of each type of assessment, and the characteristics and procedures required to ensure the effective implementation of each type of assessment.

II. Benchmark and Formative Assessments

We begin by considering benchmark and formative assessments, both of which play a major role in guiding instruction in standards-based education. Benchmark and formative tests are assessments that measure the achievement of standards. Benchmark assessments typically cover a broad range of standards. For example, a benchmark test may be used to measure the achievement of standards targeted for instruction over an extended time period. Formative assessments are typically short informal assessments covering a small number of standards.

A. Purposes of Benchmark Assessments

The central purpose of benchmark assessments is to provide information to guide instruction (Bergan, Bergan, & Burnham, 2009). A benchmark is a standard against which performance may be judged. Benchmark assessments indicate the extent to which students have mastered standards targeted for instruction. Information regarding standards mastery is used to guide the next instructional steps to promote mastery.

In accordance with current accountability requirements, standards mastery is ultimately determined by student performance on a statewide assessment. The statewide assessment provides summative information documenting standards mastery after instruction has already occurred. By contrast, benchmark assessments are administered periodically during the school year. As a result, they provide information about standards mastery while there is still time to intervene to promote learning. The periodic administration of benchmarks interspersed with formative assessments has led some investigators to refer to them as interim assessments (Perie, Marion, Gong, & Wurtzel 2007).

A second major purpose of benchmark assessments is to forecast standards mastery on the statewide test (Bergan, Bergan, & Burnham, 2009). The effectiveness of benchmark assessments in promoting standards mastery depends upon their effectiveness in forecasting statewide test performance. If a benchmark assessment indicates a failure to master standards, that indication ought to foretell poor performance on the statewide test. Likewise, if a benchmark indicated standards mastery, then mastery on the statewide test should be expected. ATI conducts research annually to evaluate the accuracy with which benchmark assessments predict student mastery of standards on the statewide end-of-year assessment. Forecasting from benchmark assessments designed collaboratively by ATI and its client partners is consistently found to yield acceptable levels of accuracy.

A third purpose of benchmark assessments is to provide information to administrators to guide instructional interventions that extend beyond individual classrooms. For example, benchmark assessments are often used to guide reteaching and enrichment interventions to promote learning. Intervention guidance requires that information on benchmark performance be aggregated beyond the classroom level. If the intervention is administered at the school level, aggregation may involve all of the classes in one or more grades. If the intervention is district-wide, aggregation may occur across all schools.

A fourth purpose of benchmark tests is to monitor progress over the course of the school year. Benchmark assessments are generally administered three or four times during the school year. Multiple benchmark assessments provide the opportunity to assess growth occurring

during the school year. Progress monitoring is necessary to ensure that growth is sufficient during the year to meet state standards often assessed toward the end of the year.

Progress monitoring may be implemented using Item Response Theory (IRT) to place benchmark assessment scores on a common scale. When assessment scores are on a common scale, progress can be measured without repeating the same items from one test to the next. The problem of teaching to the test is effectively addressed because the item sets for each of the benchmark tests are different.

B. Purposes of Formative Assessments

Formative assessments, like benchmarks, are designed to provide information to guide instruction. However, the role of formative assessment related to instruction has evolved over the years. The term formative assessment was introduced by Michael Scriven (1967) to refer to assessments providing evidence to improve curriculum. Benjamin Bloom and his colleagues modified the meaning of the term to include evaluation not only to improve curriculum, but also to improve instruction and student learning (Bloom, Hastings, and Madaus, 1971). The emphasis on improving instruction and learning remains a key component of the definition of formative assessment.

Current conceptions of formative assessment often assume an extremely close link between assessment and instruction. Formative assessment is now frequently described as an assessment process occurring during instruction that provides feedback to adjust ongoing teaching to improve the achievement of instructional outcomes. Assessment is embedded in the learning activity and linked directly to ongoing instruction (e.g., Perie, Marion, Gong, & Wurtzel 2007).

Formative assessment occurring as part of the instructional process is typically conceptualized as a classroom activity. There is generally no concern for aggregating data beyond the classroom level since assessments tend not to be common across classes. Although the term formative assessment generally refers to assessments implemented at the classroom level, there are exceptions to this rule. For example, in some cases, interventions involving formative assessment may be planned at a school or district level. School-wide or district-wide interventions may include multiple formative assessments implemented in common across classes.

In summary, formative assessment may include minute-by-minute teaching adjustments based on assessments that are a part of the instructional process. It also may include quizzes designed to measure the mastery of standards currently targeted for instruction in the classroom. Finally, formative assessment may include short common assessments directed at specific standards selected for instruction in a large-scale intervention.

C. Benchmark Test Characteristics and Assessment Procedures

Benchmark assessments are typically formal assessments that must meet rigorous standards in terms of psychometric properties and testing procedures. Psychometric studies typically include the application of IRT techniques to place test scores on a common scale. In addition, psychometric analyses must provide evidence of both reliability and validity (Bergan, Bergan, & Burnham, 2009). Validity evidence is required to demonstrate that benchmark assessments and statewide test performance are related. Reliability evidence is a prerequisite

to validity. An unreliable test is an assessment that is not related even to itself. A test that is not correlated with itself cannot be expected to correlate with other criterion measures.

Benchmark assessments are generally school- or district-wide assessments. As a consequence, these assessments often must be scheduled for large numbers of students. Moreover, the testing window must be tightly controlled to ensure that timely results are available to guide instruction. Scheduling assessments for large numbers of students is a complicated task. Bulk scheduling technology can assist in scheduling large-scale assessments.

Benchmark assessments pose security requirements affecting test scheduling and test administration. Test security is necessary to ensure that assessment results provide an accurate estimate of student proficiency. Accuracy is essential to ensure the quality of information used to inform instruction and to forecast statewide test performance. Recommended security measures and procedures for proctoring assessments are available from ATI.

D. Formative Test Characteristics and Assessment Procedures

Formative assessments are short, informal assessments typically constructed at the classroom level. Psychometric studies are not typically conducted on formative assessments. Thus, information on the psychometric properties of formative assessments is generally not available. Test scheduling, test administration, and test security are usually controlled by the teacher. The number of students taking a formative test is typically small. As a consequence, assessment procedures tend to be relatively easy to implement.

When formative assessments are implemented as part of a large-scale intervention, test procedures increase in complexity. Scheduling, administration, and security procedures analogous to those used with benchmark assessments should be adopted.

Formative assessment is widely advocated based on longstanding research supporting the assumption that it has a beneficial effect on learning outcomes (see, for example, Bergan, Sladeczek, Schwarz, & Smith, 1991; Black & William, 1998). ATI documents the implementation of short formative assessments. This makes it possible to assess the use and effectiveness of formative assessment at the district level.

III. Interim and End-of-Course Examinations

We turn next to the oldest and most familiar form of assessment in education, the course examination. Course examinations typically come in two forms, interim examinations and end-of-course examinations. Interim and end-of-course examinations are designed to assess knowledge and skills targeted in course instruction. Interim examinations assess knowledge and skills covered during part of a course. For example, an instructor may decide to administer a midterm examination half way through a course. The midterm examination will cover material addressed in the first half of the course. The end-of-course or final examination may cover material reflecting the entire course, or it may cover material presented since a previous interim examination.

A. Purposes of Course Examinations

Interim and end-of-course examinations serve two purposes. One purpose is to assess the knowledge and skills that students have acquired during participation in a course of study. Interim and end-of-course assessments hold the student accountable for learning what has been taught. As students advance through the grades, teachers place increasing responsibility for learning on student shoulders. The teacher's responsibility is to provide opportunities to learn the material presented. The student is responsible for learning the material. Interim and end-of-course examinations provide information on the extent to which students have effectively met their responsibility for learning.

The second purpose of interim and end-of-course examinations is to provide information to guide instruction. Historically, instructional guidance was limited in the main to review sessions occurring in preparation for an upcoming examination and to debriefing sessions occurring following an examination. Under the standards-based education movement, the instructional guidance function of course examinations has begun to change. More specifically, course assessments are now being designed to conform to standards-based assessment practices. For example, course content may be aligned to standards, and students may be told in advance what standards the exam will cover. Reteaching and enrichment sessions focusing on specific standards may be scheduled following an examination. In some cases, tutors may be retained to guide reteaching instruction toward mastery of standards.

The adoption of standards-based assessment practices for course examinations adds an important dimension to standards-based education. Under the No Child Left Behind (NCLB) Act, the standards-based approach focused on education in grades three through 10. The adoption of standards-based practice related to course examinations contributes to the goal of expanding education reform through high school and beyond. The desired result is a coherent educational system designed to elevate student achievement in ways that will support the development of a highly educated citizenry capable of competing effectively in the global economy of the 21st century.

B. Course Examination Characteristics and Assessment Procedures

Typically interim and end-of-course examinations are informal examinations designed and constructed at the classroom level. Test characteristics and assessment procedures are controlled in the main by the teacher who is responsible for constructing, administering, scoring and grading the test. The number of students taking a course examination is often too small to support data analyses yielding evidence of test reliability and validity. Thus, test results should be interpreted with caution.

If the examination is designed in accordance with standards-based practices, test content should be aligned to the standards targeted for instruction in the course. Test results should be used not only to grade student performance, but also to guide instruction. For example, test results may be used to inform reteaching and enrichment activities. Results might also be used to inform curriculum development and instructional practices.

In some cases, course examinations may be school-wide or district-wide assessments. In those instances, the assessment process requires a coordinated effort involving many individuals. To minimize demands on district staff, ATI provides technology to support broad-scale test scheduling, administration, scoring, and reporting of locally created interim and end-of-course examinations. Assessment Technology Incorporated's (ATI) Automated Scoring Key,

or *ASK Technology,* plays a particularly important role because it allows for automated importation of interim and end-of-course examinations built outside the system and for automated scoring and data aggregation for data from those examinations. In those instances in which course examinations are administered to large numbers of students, *ASK Technology* also makes it possible to place scores from course examinations on a common scale with other types of assessments such as benchmark assessments, and pretests and posttests. Placing scores on a common scale facilitates comparisons between performance on course examinations and other types of assessment.

To support test security, the testing schedule should minimize the window during which the test is available to examinees. In addition, procedures should be in place for controlling the availability of test materials. Procedures for addressing security issues are available from ATI.

Typically, data analyses providing evidence of reliability and validity can be conducted for school-wide or district-wide assessments. Reliability is a direct function of test length. Consequently, examinations should be of sufficient length to ensure adequate levels of reliability. Validity may be addressed in a number of ways. When a statewide assessment is available, establishing the relationship between the statewide test and the course examination is often useful because it indicates that what has been learned in the course is related to shared educational goals.

IV. Pretests and Posttests

Pretests and posttests are among the most familiar forms of assessment in education. As its name implies, a pretest is an examination given prior to the onset of instruction. By contrast, a posttest measures proficiency following instruction.

A. Purposes of Pretests and Posttests

Pretests and posttests may serve a number of useful purposes. These include determining student proficiency before or after instruction, measuring student progress during a specified period of instruction, and comparing the performance of different groups of students before and after instruction. The information derived from a pretest-posttest comparison can also be used for a variety of purposes. For example, the results of a pretest and posttest can be used by a teacher to evaluate progress towards standards mastery for his/her students. In addition, when appropriate statistical analyses are performed for data obtained from a pretest-posttest pair, administrators can also obtain results that can be used as part of an evaluation of the instructional effectiveness of the teacher. The design, analysis, and characteristics of pretest-posttest pairs designed for this purpose are described further in Section V which discusses instructional effectiveness assessment.

i. Proficiency Before or After Instruction

A pretest may be administered without a posttest to determine the initial level of proficiency attained by students prior to the beginning of instruction. Information on initial proficiency may be used to guide early instructional planning. For example, initial proficiency may indicate the capabilities that need special emphasis to promote learning during the early part of the school year. A posttest may be administered without a pretest to determine proficiency following instruction. For example, statewide assessments are typically administered toward the end of the school year to determine student proficiency for the year.

ii. Measuring Progress

A pretest accompanied by a posttest can support the measurement of progress from the beginning of an instructional period to the end of the period. For example, teachers may use information on progress during the school year to determine whether or not proficiency is advancing rapidly enough to support the assumption that students will meet the standard on the upcoming statewide assessment.

If the pretest and posttest are to be used to measure progress, then it is often useful to place pretest scores on a common scale with the posttest scores. As indicated earlier, when assessment scores are on a common scale, progress can be assessed with posttest items that differ from the items on the pretest. ATI uses IRT to place scores from pretests, posttests, and other forms of assessment on a common scale.

iii. Comparing Groups

A pretest may be given to support adjustments needed to make comparisons among groups with respect to subsequent instructional outcomes measured by performance on a posttest. The need for adjustments arises when there may be initial differences between the groups to be compared. For example, consider the familiar situation in which there is an interest in comparing achievement outcomes in a number of intervention groups to identify successful instructional practices. Differences in posttest scores between the groups could arise from group differences that were present before instruction was initiated. Pretests can be used to adjust for differences of this kind.

Group comparisons may be implemented for a number of reasons. For example, group comparisons are generally required in experimental studies. In the prototypical experiment, students are assigned at random to different experimental conditions. Learning outcomes for each of the conditions are then compared. Group comparisons may also occur in instances in which there is an interest in identifying highly successful groups or groups needing additional resources. For example, group comparisons may be made to determine the extent to which instruction is effective in meeting the needs of NCLB subgroups. Group comparisons may also be initiated to identify highly successful classes or schools. Finally, group comparisons involving students assigned to different teachers or administrators may be made if a district is implementing an instructional effectiveness initiative or a performance-based pay initiative in which student outcomes play a role in determining staff evaluations or compensation.

B. Pretest/Posttest Characteristics and Assessment Procedures

Determining how best to design and implement pretests and posttests is complex because these forms of assessment can be used effectively in many ways. The design of pretests and posttests should be informed by the purposes that the assessments are intended to serve. For example, if the pretest is intended to identify enabling skills that the student possesses that are likely to assist in the mastery of instructional content to be covered during the current year, then the pretest should include skills taught previously that are likely to be helpful in promoting future learning during the year. Similarly, if a posttest is intended to provide a broad overview of the capabilities taught during the school year, then the assessment should cover the full range of objectives covered during that period. For instance, a district-level posttest might be designed to cover the full range of objectives addressed in the state blueprint.

This section discusses selected design considerations for pretest-posttest pairs used to support group comparisons. A more in-depth discussion of design considerations for pretest-posttest pairs that will produce data for instructional effectiveness assessment is provided in Section V.

If a pretest and posttest are used to support comparisons among groups, a number of factors related to test design, test scheduling, and test security must be considered. Central concerns related to design involve content coverage and test difficulty. Both the pretest and the posttest should cover the content areas targeted for instruction. For example, if a particular set of objectives is covered on the pretest, then those objectives should also be addressed on the posttest. Targeted content increases the likelihood that the assessments will be sensitive to the effects of instruction. Both the pretest and the posttest should include a broad range of items varying in difficulty. Moreover, when the posttest follows the pretest by several months, the overall difficulty of the posttest generally should be higher than the difficulty of the pretest. Variation in difficulty increases the likelihood that instructional effects will be detected. For example, if both the pretest and the posttest are very easy, the likelihood of detecting effects will be reduced. In the extreme case in which all students receive a perfect score on each test, there will be no difference among the groups being compared.

When group comparisons are of interest, care should be taken to ensure that the pretest is administered at approximately the same time in all groups. Likewise the posttest should be administered at approximately the same time in all groups. Opportunity to learn affects the amount learned. When the time between assessments varies among groups, group comparisons may be spuriously affected by temporal factors.

Test security assumes special importance when group comparisons are made. Security is particularly important when comparisons involve high-stakes decisions. Security requires controlled access to tests and test items. Secure tests generally should not be accessible either before or after the time during which the assessment is scheduled. Galileo® K-12 Online includes security features that restrict access to items on tests requiring high levels of security. Security imposes a number of requirements related to the handling of tests. When an assessment is administered online, the testing window should be as brief as possible. After the window is closed, students who have completed the test should not have the opportunity to log back into the testing environment and change their answers. Special provisions must be made for students who have missed the initial testing window and are taking the test during a subsequent period. When a test is administered offline, testing materials should be printed as close to the scheduled period for taking the assessment as possible. Materials available prior to the time scheduled for administration should be stored in a secure location. After testing, materials should either be stored in a secure location or destroyed.

V. Instructional Effectiveness Assessment of Student Progress

Instructional effectiveness (IE) assessment of student progress is designed to support value-added modeling and other statistical techniques known to be useful in the identification of critical factors affecting student achievement. The design of IE assessments should support data analysis initiatives aimed at pinpointing critical instructional influences on student learning. School and teacher effects on student academic progress are among the most important of these influences.

A. Purposes and Analyses of Instructional Effectiveness Assessments

Value-added analyses or other types of analyses of IE assessments of student progress may be used to inform policy decisions regarding resource allocations aimed at elevating student achievement. Analyses of IE assessments may also be used along with other indicators to evaluate the performance of teachers and administrators. Such evaluations may be used to inform professional development initiatives aimed at improving the quality of instruction.

Analyses of IE assessments of student progress require the implementation of complex mathematical models. A number of different modeling approaches are available. ATI uses two approaches. The first is value-added modeling implemented using Hierarchical Linear Modeling (HLM) (Raudenbush & Bryk, 2002). Value-added modeling provides a highly flexible approach and makes it possible to identify sources of variability in student performance associated with the schools and classes to which students are assigned. For example, the overall school culture may encourage or inhibit student learning. This school-level influence on student academic progress can be identified using value-added models so that differences between schools can be identified. Value-added modeling also enables the identification of classroom-level contributions to student progress that are independent of the school-level influences. Student, class, and school background variables may also be included in value-added models. For example, variables defining student subgroups may be represented. In addition, intervention effects may be addressed using value-added modeling.

The second approach to the analysis of IE assessments of student progress involves the use of a form of categorical data analysis termed Categorical Growth Analysis (CGA) . CGA has many of the benefits of value-added modeling. CGA can identify school, and teacher effects. It can assess effects on learning for subgroups, and it can be used to assess student progress. ATI's implementation of CGA provides three additional benefits:

- ATI CGA relates the evaluation of instruction directly to the achievement of AYP measurable objectives for subjects covered on statewide assessments.
- ATI CGA assesses increases in the probability of standards mastery associated with the teacher and school to which the student is assigned. The focus is on the achievement of goals reflecting student progress.
- ATI CGA can be used with relatively small data samples. Thus, CGA is well suited for use in local instructional effectiveness initiatives.

Further details regarding the value-added modeling and CGA analyses offered by ATI can be found in Bergan et al. (2011).

B. Instructional Effectiveness Assessment Test Characteristics and Assessment Procedures

Much of the benefit derived from IE assessment is due to the nature of the analysis techniques that are applied to the student performance data. However, the assessments themselves must be designed properly in order to facilitate such analyses. IE assessments are generally administered in pairs: a pretest and a posttest. ATI offers pre-made instructional effectiveness pretests and posttests in a wide variety of state-tested grades and content areas. Participants in ATI's *Community Assessment and Item Banking (CAIB) Initiative* also collaborate with ATI and other districts to develop items and assessments appropriate for IE assessments in non-state-tested areas. Participants gain access to all items and assessments developed as part of the initiative. A number of assessments and more than 8,000 items have already been

developed representing a broad range of subjects including social studies, art, music, and foreign language.

Ideally, the IE pretest is administered at the beginning of the school year, before instruction has begun, and the IE posttest is administered at the end of the school year at the conclusion of instruction. To the extent possible, all students in the district should be tested at the same time for both the IE pretest and the IE posttest. For large districts this may require a testing window of up to two weeks, but the testing period should be as brief as possible to ensure that the snapshots of student ability that are provided by the assessments fairly capture the students in all classrooms at the same moment in terms of their educational development. The two tests should be comprehensive and should address the same content, but they should contain different items in order to avoid concerns that observed student gains are due to teaching to the test rather than actual student learning. The scores from the IE pretest and IE posttest are placed on a common scale via IRT techniques so that changes in student achievement levels can be identified. One of the previously described statistical analyses is then applied to the student performance data to identify the degree to which the teacher or school had a positive impact on student achievement.

Given that the data provided by IE assessment may be used in making important decisions, it goes without saying that the assessments should be both reliable and valid indicators of student performance. Reliability and validity are particularly important considerations when IE assessment is used in staff performance evaluations. Staff evaluations based on student performance on unreliable or invalid assessments cannot be justified. Inclusion of data from such assessments puts both evaluated staff and administrative evaluators at risk.

A reliability coefficient of .90 or higher provides an acceptable level of internal consistency for IE assessments. Reliability coefficients in the .80s may be justified when multiple measures of instructional effectiveness are used. Reliability is a direct function of test length. ATI research indicates that tests of 70 items or more are likely to yield reliability coefficients in the .90s. Tests of approximately 45-50 items can be expected to yield coefficients in the .80s. ATI's IE pretests and posttests typically contain approximately 45 items.

Even when assessments are of adequate length, reliability may be compromised in circumstances in which student knowledge of the content being assessed is extremely low. This may occur on a pretest introducing entirely new content. To avoid compromising reliability, it is recommended that pretests include some content to which students have been previously exposed. For example, a pretest for fifth-grade math might include some items covering fourth-grade content as well as fifth-grade items of appropriate difficulty for students at the beginning of fifth grade. ATI's IE Pretests include prior-grade-level content to help ensure the reliability of the assessments. IE Posttests contain only current-grade-level content.

Validity may be assessed by examining the relationship between each IE assessment and a criterion measure such as a statewide test. ATI research indicates that on average validity coefficients of approximately .75 can be expected.

Test security is particularly important with regard to IE assessment. If the student scores on the assessment are to be used to inform teacher evaluations, it would not be appropriate for teachers to have access to the assessments before they are administered. For example, the teacher review process that is in place for district benchmark assessments is not appropriate for

IE assessments. In addition, the assessments must be housed in locations, both physical and electronic, to which teachers and evaluated administrators do not have access.

Although IE assessments must be highly secure before and during their administration, it may be appropriate to allow teachers to access the assessments after they have been administered. An IE pretest administered at the beginning of the school year can be very useful in guiding instruction. Given the classroom time that will be dedicated to administering an IE assessment that is long enough to ensure good reliability, it may be considered wasteful to prevent teachers from using the data to guide instruction. District administrators may decide that, once the assessment has been fully administered district-wide, teachers may be permitted to view reports of the student data. Many reports can provide useful, formative data without providing access to the items themselves. For example, the Galileo® K-12 Online *Intervention Alert, Intervention Planning*, and *Development Profile* reports all indicate the number of students who have demonstrated mastery of each standard on an assessment as well as the standards on which each student has demonstrated mastery. This information is available without the need to see the assessment items themselves and can be useful in assessing the readiness levels of the students at the beginning of the year. Since, for reliability purposes, the IE pretest contains content from the state standards at the prior grade level, they can provide the teacher with an indication of any skills from the prior grade level that students may be lacking. By limiting teacher access to reports that do not display the assessment items, test security can be preserved and the student performance data can be used formatively. This compromise would give teachers the knowledge of what content is included on the IE pretests and posttests in terms of which state standards are included, but since the assessments are known to be comprehensive, knowledge of the test blueprint would not provide the teachers with much information they did not already have. Even so, in order to preserve the integrity of the test it would be crucial to "close" the assessment so that no more student data could be recorded after the reports were made available to teachers.

With regard to the actual administration of IE assessments, the conditions should be as similar across classrooms as possible. Additionally, care should be taken to ensure that each student score is properly linked to the appropriate classroom. For example, some districts schedule students for benchmark assessments in a blanket manner, scheduling each assessment for every class in the school so that students may take the assessment in whichever one of their classes is most convenient. While this practice may simplify the scheduling process, it makes it virtually impossible to link a student's assessment data to the relevant class and, ultimately, to the teacher that is responsible for student learning in the relevant content area. Such a scheduling practice would render the assessment data useless for evaluating instructional effectiveness and providing data for teacher evaluations. As part of scheduling an assessment, ATI enables the user to designate a responsible teacher for each class. The responsible teacher is then directly associated with the group of students in that class for purposes of analyses and reports related to IE assessment.

VI. Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) is a form of assessment in which the items administered are adapted to the ability levels of the students. The goal of adaptation is to increase the precision of measurement. Precision is maximized when the probability of a correct response by a student is exactly .5 (USDE, NCES, 2003). CAT algorithms are designed to select items for students of a given ability level yielding probabilities as close to the .5 criterion as possible. Prior information about student ability is used as a basis of item selection. Ability

estimates obtained from previously administered items guide the selection of subsequent items. CAT provides high levels of efficiency in the assessment of student ability. Given a large item pool comprised of highly discriminating items, CAT administration of small numbers of items can yield test reliabilities equivalent to those for a much longer non-adaptive test (Wainer, 2000; USDE, NCES, 2003).

Adaptive assessments fall into two broad categories (Wainer, 2000). These may be labeled item-by-item adaptive assessment and multi-stage adaptive assessment. In the item-by-item approach, adaptation occurs following a student's response to each item. The number of items administered to any given student may vary. If the student responds correctly to an item, a more challenging item will be presented next. If the student fails the item, a less challenging item will be administered. Testing continues until a pre-established level of precision is achieved.

In the multi-stage approach, adaptation occurs following administration of a set of items. The number of items administered to a given student is fixed. The student responds to an initial item set designed to route the student to a subsequent set appropriate to his or her ability. If the student performs at a high level, he or she will be routed to a challenging set of items. If the student initially performs at a relatively low level, he or she will be routed to a less demanding set of items.

The ability to control item exposure is a significant consideration in determining which form of CAT to select (Edwards & Thissen, 2007). CAT assessments are often administered to individual students over an extended time span. When the same items are used repeatedly, test security may be compromised (Wainer, 2000). Exposure control provides a useful way to ensure security. Controlling item exposure can be a highly complex challenge when item-by-item CAT is implemented. The complexity of exposure control techniques and the computational demands necessary to achieve control have led many test developers to opt for the multi-stage approach to address the exposure control issue (Edwards & Thissen, 2007).

ATI automates the construction of adaptive tests. Tests are generated from an *Assessment Planner* that defines the item pool to be used in selecting items for the adaptive assessment. Automated construction allows the district or school to construct customized adaptive tests to meet unique local needs. For example, a district may construct an adaptive test to be used in determining placement in a locally designed advanced algebra course. Automated construction increases the testing options available for adaptive testing. In addition, it supports accommodations to continually changing standards which are a hallmark of contemporary education.

A. Purposes of CAT

Adaptive assessments serve many of the purposes outlined for the various types of assessment discussed in this document. In some cases, adaptive assessments offer a clear advantage over non-adaptive measures. In other instances, non-adaptive approaches may be preferred. The discussion that follows summarizes the major uses of the adaptive approach.

i. Guiding Instruction

CAT is not explicitly designed to provide information to guide instruction. However, CAT has been widely used for that purpose. In order for CAT to be used effectively to guide instruction, it is necessary to put in place CAT item selection rules that choose items based on their alignment to standards. Selection rules involving alignment to standards are needed to

ensure that test content adequately reflects standards targeted for instruction. These rules must work in concert with rules that maximize measurement precision.

It is easier to establish usable standards-alignment rules for multi-stage adaptive assessments than for single-stage instruments. Item selection for item sets included in a multi-stage adaptive assessment is completed prior to the time that students are assessed. Moreover, as previously mentioned, variations in item exposure can be much more easily controlled in a multi-stage assessment than in an item-by-item adaptive assessment (Edwards & Thissen, 2007).

The decision to use or not use adaptive assessment to guide instruction will depend on the specific goals to be accomplished through the assessment initiative. If measuring the mastery of specific standards is the major goal of assessment, a standard benchmark or formative assessment may be preferred over an adaptive test. However, if efficient measurement is a major concern, then CAT may be the preferred option. However, other factors must also be considered. For example, CAT may not be practical if the necessary technology to support adaptive testing is not available.

ii. Monitoring Progress

CAT is well suited to the task of monitoring progress. CAT assessments are designed to place scores from multiple assessments on a common scale, which is critical to progress monitoring. IRT techniques provide the needed technology to achieve the goal of establishing a common scale. CAT also increases the number of standards that may be covered in an assessment while also minimizing the number of items administered to any given student. Given the measurement efficiency provided by CAT, it is often an attractive option for monitoring progress. However, if the goal is to monitor progress related to specific standards targeted for instruction at different points in time, a series of benchmark assessments may be preferred.

iii. Screening

CAT is well suited to the task of screening. CAT assessments can provide efficient measures of ability with high levels of precision. CAT is especially useful in cases in which there is a need to set a cut score near an extreme point in the score distribution. Multi-stage adaptive assessments can be designed to maximize precision at specific points in the ability distribution. For example, a cut score for an RTI screening might be established toward the lower end of the score distribution. This would ensure maximal precision at the point where intervention decisions are needed.

iv. Placement

CAT assessments are frequently used for placement purposes. For example, a CAT assessment may be used to determine grade placement for a newly enrolled student. CAT assessments are also widely used in making school admission decisions. The Graduate Record Exam offered by the Educational Testing Service is a prominent example. Students across the nation take this examination and make the results available to graduate schools as part of the process of applying for admission to graduate studies.

The efficiency of a CAT assessment makes it an attractive option for placement purposes. However, if the placement decision requires information on the mastery of specific

standards as may be the case in advanced class placement, a non-adaptive placement test may deserve consideration.

B. CAT Characteristics and Assessment Procedures

CATs have a number of characteristics that are important to consider in planning an assessment program. First, as the name Computerized Adaptive Testing indicates, a CAT is not intended for offline administration. The earliest forms of adaptive assessment were administered offline. Moreover, there are a number of offline adaptive assessment instruments currently in use. Individual assessment batteries that route examinees to different items based on their initial performance provide familiar examples. Nonetheless, offline administration for adaptive assessments is generally not practical. The time-saving features that are an important component of the attractiveness of CAT can be severely compromised in offline administration.

A second CAT characteristic that requires consideration in assessment planning involves the types of items that may be included in a CAT assessment. CAT assessments are designed for items that are inexpensive to construct and that can be scored automatically. Expense is a significant issue because CAT generally requires support from large item banks. CAT is not well suited to handle open response items that are expensive to construct and that are scored manually (Wainer, 2000). For example, CAT is not the preferred option for essay exams scored manually using a rubric. CAT can be used when only a portion of the assessment lends itself to automated scoring. For instance, open response items can be included as a separate item set scored manually following the adaptive portion of an assessment.

Assessment procedures associated with Computerized Adaptive Testing are markedly different than those associated with non-adaptive testing. One major difference involves scheduling flexibility. Non-adaptive tests are generally administered within a fixed time period. CATs are often administered within a highly flexible time frame. For example, individual students may be scheduled to take a CAT at any of several available times during the day. Moreover, in some cases test availability may extend over a number of months.

Flexible scheduling is accompanied by reduced security requirements. For example, proctoring requirements are reduced because students are not exposed to an identical set of items. Management of test materials is also simplified because of the differences in test content occurring across students.

Flexible scheduling may also reduce equipment requirements. When students are scheduled individually at different times during the day and different days during the week, the number of computers required to conduct the assessment may be reduced.

VII. Testing Arrays

In some cases it is useful to construct assessments from testing arrays comprised of multiple sets of items. In a testing array, each item set is typically composed of a group of items aligned to a particular standard or a small number of standards. For example, a testing array designed to assess early reading proficiency might include item sets for standards addressing letter naming, phonemic awareness, phonics, vocabulary, fluency, and other capabilities. Each item set in a testing array may be scheduled, administered, and scored independently. In addition, scores may be obtained for groups of item sets, and for all of the item sets in the array. For example, IRT could be used to generate a Developmental Level (DL) score for the letter

naming, phonemic awareness, and phonics item sets in the array. A DL score might also be generated for all of the item sets comprising the early reading array. In ATI's comprehensive system, testing arrays may be constructed by system users. For example, a user might select a set of short formative assessments from the *Array Builder* interface to form an array. The ability to create custom testing arrays enhances the system's capability to meet local assessment needs.

A. Purposes of Testing Arrays

A testing array provides a reliable and valid assessment utilizing item sets administered over multiple occasions. The assessment can be used for the purpose of guiding instruction and/or documenting learning outcomes.

There are a number of circumstances that call for the testing array approach. One of these involves the assessment of young children. Testing arrays reduce the amount of time required for assessment on any given occasion. This can be beneficial in assessing young children with short attention spans. Testing arrays may be particularly helpful when assessments are administered to one child at a time because the amount of time that any child is removed from instructional activities is minimized.

Testing arrays are often beneficial in assessing intervention outcomes. For example, consider the situation in which a district-wide intervention is initiated to increase student mastery of standards involving knowledge and skills related to fractions. Suppose that the intervention was implemented over a number of weeks and that a formative assessment was administered each week to determine mastery of standards covered during the week. Implementation of a testing array would combine the formative assessments to provide an overall assessment of sufficient length to provide a reliable assessment of student knowledge and skills for the entire intervention. Information of this kind could contribute significantly to the evaluation of intervention effectiveness.

B. Testing Array Test Characteristics and Assessment Procedures

Testing arrays have a number of unique characteristics. Each item set in a testing array may be treated as independent assessment. In addition, each item set is assigned to membership in the testing array. Psychometric properties may be established for each item set, for selected item sets, and for the full array.

Scheduling requires a date and a time for each item set to occur within an overall testing window defining the period of time during which the scheduled assessments must be completed. Security and proctoring should follow procedures analogous to those used for benchmark assessments.

VIII. Dialogic Assessment

Assessment in the form of teacher questions has been a cornerstone of dialogic teaching since the appearance of the approach in Socratic dialogs more than two-thousand years ago. Now ATI's dialogic assessments take the form of online and offline assessments that are an integral part of an instructional dialog. Dialogic assessments may include informal questions arising during classroom interaction. They also may be comprised of assignments including assessments interspersed at critical points in the instructional process.

A. Purposes of Dialogic Assessments

The fundamental purpose of dialogic assessment is to create a seamless integration between assessment and instruction, thereby making it possible to use assessment information to rapidly adjust instruction to promote learning. Dialogic assessment information indicates the extent to which standards targeted for instruction in a dialog are being mastered during the instructional process. This information can be used during the course of the dialog to plan next instructional steps. It also can be used to plan interventions following the dialog to assist students in need of additional instruction to achieve targeted goals.

A second purpose of dialogic assessment is to support the transition from paper-and-pencil learning to digital learning. Contemporary instructional dialogs support a broad range of question types including the many innovative item types emerging as the nation moves toward digital learning. These assessment innovations often make it possible to assess capabilities not easily assessed offline. For example, computers make it easy to assess performance on timed tasks that may be difficult to assess offline. Dialogs provide a convenient way to initiate digital learning during the current period of transition in which the technological infrastructure to support digital learning is under development in many educational sites.

A third purpose of dialogic assessment is to support the assessment of capabilities that cannot be easily assessed within the time limits allotted for a typical assessment. For example, the assessment of capabilities reflecting high levels of Depth of Knowledge often require extended amounts of time that cannot be easily accommodated under the time constraints associated with a typical test (e.g. Webb, 2006). Assessment of complex problem-solving tasks provides a familiar example. Dialogic assessment offers an efficient way to assess performance on tasks of this kind.

B. Dialogic Assessment Characteristics and Procedures

The central characteristic of a dialogic assessment is that it is administered during and immediately following the completion of an *Instructional Dialog*. The assessment process typically involves a number of assessments spread across an extended time span. The overall time necessary to complete the assessment is relatively long. Yet, the time allotted for any particular assessment is very short. The administration of multiple assessments over time creates scheduling flexibility that can be beneficial in a number of ways. For example, in those cases in which the number of computers available for online administration is limited, small numbers of students may be scheduled to take the assessment in each of a number of testing sessions.

A second important characteristic of dialogic assessments involves the manner in which they are planned and administered. Dialogic assessments may be planned, constructed, and administered by teachers for use in their classrooms. However, they may also be planned using the *Assessment Planner* under the guidance of ATI's Educational Management Services Department. This would be an appropriate option in those instances in which the Dialog is planned for district-wide or school-wide implementation. For example, if the district has planned a district-wide benchmark assessment aligned to standards in the district pacing calendar, planning and administering one or more district-wide Dialogs aligned to the pacing calendar standards would support instruction designed to prepare students for the upcoming benchmark. Since Dialogs can be modified by users, the manner in which the standards were taught could vary from class to class. Such variation is important to accommodate the individual learning

needs of students. Nonetheless, the standards covered during instruction in all classes would likely include those specified in the pacing calendar.

A third important characteristic of dialogic assessment is that it provides a permanent record of what has been taught as well as what has been learned. Research on the opportunity to learn has shown that information regarding what has been taught plays a particularly important role in determining learning outcomes. Dialogic assessment provides the information needed to assess the association between what has been taught and what has been learned. This kind of information is essential to determine the effectiveness of instruction.

A fourth important characteristic of dialogic assessment involves its potential to positively impact test reliability and validity. Since dialogic assessments are designed to occur during instruction, they may often be administered in close proximity to other assessments. For example, a dialogic assessment may occur during a period leading up to the administration of a benchmark assessment. In circumstances of this kind, it may be useful to use ATI's *Test Array* feature to combine the dialogic assessment with the benchmark assessment to produce a new and longer assessment. Since test reliability is a direct function of test length, the combination is likely to yield an assessment with greater reliability than either of the component assessments in the array. Moreover, since the larger assessment provides additional information regarding standards mastery, it could increase validity by increasing effectiveness in forecasting statewide test performance.

IX. Screening and Progress Monitoring Assessments

Screening and progress monitoring assessments are not new tools within a comprehensive assessment system; however, with the growing implementation of response-to-intervention (RTI) initiatives these tools have been accorded increased prominence. The successful use of these types of assessments requires an understanding of the purposes and defining characteristics of each type as well as important design considerations, and associated testing procedures.

A. Purposes and Defining Characteristics of Screening Assessments

The main purpose of a screening assessment is to quickly and efficiently identify a set of students who are likely to experience a certain outcome in the future. A screening assessment accomplishes this task by assessing student performance on a set of critical skills that are linked to the future outcome of interest. A screening assessment can be designed to predict either positive or negative outcomes. For example, with respect to predicting positive outcomes, a screening assessment might be used to assess academic readiness by identifying preschool students who perform well on critical skills that are linked to success in kindergarten. Conversely, with respect to predicting negative outcomes, a screening assessment might be used to assess the risk of future reading difficulties by identifying students who perform poorly on a set of critical skills that are linked to success in reading.

The major benefit of a well-designed screening assessment is that it accurately identifies students who are likely to experience a certain outcome at an early stage when the information can be used to guide decision-making. In particular, identifying students at risk for a negative outcome provides an opportunity to intervene at the earliest point possible so that the negative outcome can be prevented. Although a screening assessment focused on a small set of critical skills is highly useful as a method of quickly identifying at-risk students, it does not provide

enough information to guide targeted intervention. Typically, additional diagnostic assessments will be administered to students identified as at risk to determine the exact nature of the student's deficiencies. In addition, progress monitoring assessments will be administered to evaluate the student's progress over time in response to instruction or targeted interventions.

Universal (i.e., school-wide or district-wide) screening plays an important role in RTI initiatives. Within an RTI initiative, screening assessments typically are administered to every student one or more times throughout the year. Research suggests that many students identified as at risk by one-time screening make adequate progress without targeted intervention (Compton, Fuchs, Fuchs, & Bryant, 2006). For this reason, many RTI models recommend screening at least three times throughout the year and administering progress monitoring assessments in the interim to assess student progress in response to classroom instruction and/or interventions. When used in this manner, screening can serve not only the primary purpose of identifying at-risk students, but also the secondary purpose of evaluating the adequacy of curriculum and instructional practices. If a large number of students within a class or school are consistently identified as at risk by screening assessments and fail to make progress across a series of progress monitoring assessments, it suggests that the curriculum and/or instructional practices in that class or school should be revised.

B. Design Considerations for Screening Assessments

Since screening assessments are designed to be quick and efficient, they are often brief and easy to administer. Screening assessments are usually not comprehensive in terms of content; rather, they focus on a small set of critical skills linked to the outcome of interest. Perhaps most critically, screening assessments must do a good job of distinguishing students who will experience the outcome of interest from those who will not.

Designing a successful screening assessment entails careful consideration of a number of factors. The first task is to carefully define the outcome to be predicted and the indicator measure for that outcome. For example, a screening assessment might be designed to predict student performance on a standards-based statewide reading assessment administered at the end of third grade (i.e., the indicator measure). A negative outcome might be defined broadly as failing to show mastery on the statewide assessment or more specifically such as falling into the lowest performance category or falling below some specific percentile rank. The details of the way the outcome is characterized will have significant consequences for the number of students who will be identified by the screening assessment, so practical considerations such as available resources may also be taken into consideration. In the context of a response-to-intervention initiative, the focus of screening assessments is typically on identifying low-achieving students at high risk for negative outcomes; however, screening assessments can also be designed to identify high-achieving children for such purposes as placement in a gifted program.

After the predicted outcome and indicator measures have been identified, the set of skills to be assessed and the items testing those skills must be specified. To ensure a valid screening assessment, the selected skills should be linked to the outcome of interest. These links should be based on scientific research and should be continuously evaluated over time. Items or tasks selected for screening assessments should also provide maximal information in the ability range targeted by the screening assessment. For example, a difficult item that assesses an advanced reading skill will distinguish well between proficient and highly proficient students; however, this item will not distinguish well between students of lower ability. Such an item would be appropriate for a screening assessment designed to identify highly proficient

students for placement in a gifted program, but would not be appropriate for a screening assessment designed to identify students who will experience severe reading difficulties. ATI's extensive item bank contains information about the properties of items derived from Item Response Theory analyses of the responses of a large number of students of varying abilities. This information can be used during the design of a screening assessment to select items that target the desired ability range. ATI also provides districts with technology that supports the design and creation of customized screening assessments and conducts research to continuously evaluate and optimize the item- and test-level information of the assessment for the targeted population.

Once the content of the screening assessment has been selected, the next step is to set a cut score on the assessment that divides students into those who are likely to experience the targeted outcome and those who are not likely to experience the targeted outcome. Cut scores can be set in a variety of different ways. One approach employed at ATI is known as equipercentile equating. Under this approach, the cut score for the screening assessment is set so that the percentage of students who will be identified as at risk matches the percentage of students who experienced the targeted outcome on the indicator measure for the district from the previous year. This approach has been successfully applied in setting cut scores for benchmark assessments that can be used to accurately identify students at risk of failing to show mastery on the statewide end-of-year assessment.

In order to provide useful information for decision-making, it is critical that a screening assessment is accurate not only in identifying students who will experience the targeted outcome (i.e., shows adequate sensitivity) but also in identifying students who will not experience the targeted outcome (i.e., shows adequate specificity). For example, if a screening assessment designed to identify students at high risk for future reading difficulties does not have adequate sensitivity, it will fail to identify as at risk some students who will indeed experience future reading difficulties. In other words, some students will not receive the intervention that they need, which is undesirable. In contrast, if the screening assessment does not have adequate specificity, it will identify some students as at risk who will not experience future reading difficulties. In other words, some students will receive interventions that they do not need, which will tax the available resources unnecessarily.

When students will be targeted for intervention based on the results of the screening assessment, sensitivity is typically considered more important than specificity. Therefore, within reasonable limits, screening assessments are typically designed to err on the side of identifying more students for intervention than the number actually at risk. In RTI initiatives, the use of repeated screening assessments, more in-depth diagnostic assessments, and follow-up progress monitoring assessments helps to detect students who were initially incorrectly identified as at risk. Once the targeted outcome has occurred, it is also important to conduct research evaluating the sensitivity, specificity, and overall accuracy of the screening assessment and to modify the assessment or cut score as needed.

C. Purposes and Defining Characteristics of Progress Monitoring Assessments

The main purpose of a progress monitoring assessment is to assess whether a student is making adequate progress over time in response to the curriculum, instructional practices, and/or interventions to which the student is exposed. Any interim assessment that is used to evaluate student progress can be considered a progress monitoring assessment; however, this term has taken on a more specific meaning in the context of RTI initiatives. In this context, progress monitoring assessments are brief, frequently administered assessments targeting one

or more critical skills. Often the progress monitoring assessments employed in RTI initiatives assess the accuracy and speed of student responses simultaneously by limiting the time the student has to complete the assessment.

Progress monitoring assessments typically focus on an individual student rather than a classroom or school and evaluate whether the student is showing adequate growth over time. If the student fails to show adequate growth across a series of progress monitoring assessments, it suggests that the instruction or intervention the student is receiving is not sufficiently effective for that student and needs to be adjusted. If the student is currently only receiving standard classroom instruction, the student may be placed in an intervention tailored to their specific needs. Determining the content of this intervention may require more in-depth diagnostic assessment. If the student is already receiving an intervention, the intensity of the intervention may be increased, the focus shifted, or the approach modified.

The major benefit of progress monitoring assessments is that they provide an evaluation of student progress at regular intervals which can be used to support timely decision-making. Within RTI initiatives, progress monitoring assessments are typically administered every one or two weeks.

D. Design Considerations of Progress Monitoring Assessments

Progress monitoring assessments are designed to measure progress towards a goal over time. Therefore, a progress monitoring assessment must do more than simply indicate whether or not a student meets a certain standard; it must indicate whether a student is making adequate progress toward meeting that standard. For this reason, progress monitoring assessments must be capable of measuring small increments of growth. The more data an assessment collects about a student, the more capable it can be of distinguishing small increases in student ability. For example, an assessment containing five items would only potentially be able to divide students into five categories based on whether the student answered one item correctly, two items correctly, and so on. In contrast, an assessment containing 40 items would potentially be able to divide students into 40 categories based on their responses. In designing progress monitoring assessments, then, the goal is to maximize the amount of data collected about the student; however, practical considerations such as the time required to administer the assessment and the desired frequency of administration may also be considered.

The content of progress monitoring assessments must also be carefully selected. The items and tasks selected for progress monitoring assessments should be capable of revealing small increments of growth in all students who are intended to take the assessments. For example, a progress monitoring assessment that contains 20 items of relatively low difficulty will not be capable of measuring progress in a student of high ability who answers all the items correctly on the first progress monitoring assessment. If a progress monitoring assessment is intended to be administered to students whose abilities range from high to low, then the items, skills, or tasks selected for the assessment must provide information about students with a range of abilities. In contrast, if a progress monitoring assessment is intended to be administered only to students of low ability, then the items, skills, or tasks on the assessment may be targeted more specifically to those students.

Since progress monitoring assessments are designed to be administered at regular intervals to the same student, multiple forms of the assessment must be created. In order to effectively evaluate growth over time, these forms must be equated so that student scores are directly comparable across assessments. One way to achieve comparability of student scores

across assessments is to assess the same content on each form and match the difficulty of the forms. This approach is typically used to construct multiple alternate forms of a progress monitoring assessment that targets one or more critical skills that serve as indicators of overall ability. For example, a set of progress monitoring assessments evaluating letter naming ability and matched for difficulty might be constructed for administration in kindergarten. ATI's extensive item bank contains information about the difficulty of items that is based on the responses of a large number of students of varying abilities and that is continuously updated. This information can be used to construct assessments of similar difficulty where appropriate.

Another way to achieve comparability of student scores across assessments is to place the scores from these assessments on the same scale using test scaling procedures based in Item Response Theory. This approach is typically used to construct multiple progress monitoring assessments that evaluate progress towards a goal. For example, ATI uses this approach to place benchmark assessments administered throughout a school year on the same scale so that student scores on these assessments are directly comparable. In the context of progress monitoring, these procedures could be used to place scores from a set of progress monitoring assessments evaluating progress towards mastery on the statewide math assessment at the end of 10th grade. One benefit of this approach is that scores can be equated across assessments that do not contain the same content, so the progress monitoring assessments can be customized to the instruction students are receiving in the classroom. Using technology provided by ATI, districts and schools can create standards-based progress monitoring assessments customized to their curriculum and pacing guide.

The purpose of administering a set of progress monitoring assessments is to evaluate whether a student is showing adequate progress towards a goal. Therefore, once a set of progress monitoring assessments has been constructed, the next task is to establish what constitutes adequate progress. A variety of approaches can be used to establishing adequate progress over a given time period. One approach used at ATI is to define adequate progress normatively. For example, ATI conducts annual research to calculate the expected amount of monthly growth for a given content area and grade level based on the performance of a large number of students on benchmark assessments. This approach accommodates the evaluation of progress towards reaching an established standard such as mastery on the end-of-year statewide assessment and is particularly useful in initiatives designed to reduce the number of at-risk students or to achieve goals with respect to adequate yearly progress (AYP).

Another approach is to define adequate progress based on a desired goal and the targeted time period for reaching that goal. This approach accommodates the setting of individualized goals such as are expressed in individual education plans (IEPs) for special education students. Progress that maintains the desired growth trajectory is seen as adequate and suggests that the instruction and/or intervention the student is receiving is sufficiently effective. Progress that exceeds the desired growth trajectory is also adequate but suggests that the goal should be raised to challenge the student. Progress that fails to maintain the desired growth trajectory is inadequate and suggests that the instruction/intervention should be revised.

E. Screening and Progress Monitoring Assessment Procedures

Screening and progress monitoring assessments are low-stakes assessments. Therefore, these assessments do not require the extensive security procedures associated with high-stakes assessments. Typically, screening assessments are administered district-wide one or more times a year necessitating coordinated scheduling, online automated scoring, and

statistical analysis of the aggregated data. Depending on the specific implementation, progress monitoring assessments may be administered district-wide or on a much smaller scale. When progress monitoring assessments are administered to a small subset of students or even a single student, administration is typically handled either by the classroom teacher or by a resource teacher or other specialist.

As discussed elsewhere, technology provided by ATI facilitates the bulk scheduling of assessments. In addition, assessments can be administered in a variety of formats such as offline, online, or using handheld response pads. CAT technology may be particularly useful for screening and progress monitoring assessments since this technology provides a highly reliable and precise estimate of student ability in an extremely efficient manner. If CAT technology is to be used, it is important to consider during the planning and assessment creation process a variety of additional factors relating to the design and implementation of assessments using adaptive testing.

Regardless of the format of administration, all scoring can be completed online in an automated fashion. For district-wide screening assessments, ATI's Research and Development staff can also conduct statistical analyses using IRT techniques as well as research evaluating the sensitivity, specificity, and overall forecasting accuracy of these assessments. If progress monitoring assessments are administered district-wide, statistical analyses using IRT techniques may also be conducted; however, these analyses are not appropriate for assessments administered only to a subset of the population or to a single student.

The main challenge associated with the successful use of screening and progress monitoring assessments is the collection of the resulting data in a format that is useful for decision-making. ATI provides flexible reporting technology that allows users to view assessment results for the individual student, intervention group, class, school, or district as well as the capability to filter aggregated data by demographic or other student characteristics (e.g., ethnicity, ELL status). By maintaining an online history of assessment results that can be accessed by any staff member with the appropriate permissions, the technology provided by ATI can support the collaborative decision-making and problem-solving discussions required to successfully use the results of screening and progress monitoring assessment to evaluate and optimize curriculum, instructional practices, and interventions for individual students and the school or district as a whole.

X. Placement Tests

Placement tests are a particularly important form of assessment because they can have life changing consequences for students. Placement tests are used to inform decisions about the educational programs that will be made available to students. In the best of circumstances, those decisions are based on what is best for the student. However, placement assessments may also occur when resources are limited, and the goal is to select those students who are most likely to benefit from highly valued advanced educational opportunities. One of the finest attributes of the American educational system is that it provides multiple opportunities to pursue paths leading to advanced knowledge. The placement test can and should provide the best available option at the time of the assessment without limiting future options.

A. Purposes of Placement Tests

The fundamental purpose of placement tests is to provide information that can be used to inform decisions regarding the appropriate educational placement of students. The context of the placement decision may vary. One common use for placement tests is to determine whether a student is ready to take an advanced course. For example, a student that is just beginning eighth grade may take a placement test to see whether he or she is ready to take an algebra course. Another use for a placement test might be to determine the appropriate grade level into which a new student should be enrolled. While the situations vary, the primary goal of a placement test is the same: obtain a measure of student ability and content mastery in order to inform decisions about appropriate educational placement. In practice, placement decisions should and generally do consider factors in addition to the student's score on the placement test, such as past course grades or the social impact of placement in an advanced setting.

B. Placement Test Characteristics and Assessment Procedures

The design of placement tests is a highly complex process that involves unique challenges related to test content and test psychometrics. In addition, placement tests often pose significant constraints on testing procedures. The discussion that follows outlines the challenges associated with the development and implementation of placement tests and suggests ways of addressing those challenges.

i. Content of Placement Tests

Placement tests are more useful in some content areas than others. In order for the concept of placement tests to be useful, the content area must be one in which a series of courses exist in a hierarchical sequence and in which knowledge and skills obtained in the earlier courses on the sequence are prerequisites for courses later in the sequence (Whitney, 1989). Placement tests may be useful, for example, in mathematics and foreign language courses because in both cases there is a relatively clear progression of skills. However, it is more difficult to argue that one must be able to demonstrate competence in world history before taking a course in American history, and so the use of placement test in the context of history courses may not be appropriate.

In traditional education, especially at the college level, the content validity of educational placement tests relied heavily on the judgment of faculty members responsible for the instruction of the target courses (Whitney, 1989). After all, it was they who best understood the prerequisites for a given course. Today, within the context of standards-based K-12 education, the design of placement tests is greatly simplified in many cases. This is because the course of study in content areas such as mathematics is clearly laid out by state standards. For example, consider a case where the state standards outline an integrated mathematics course through the eighth grade, followed by a course of algebra in the ninth grade. If a teacher wants to know whether a beginning eighth-grade student is eligible to skip the regular eighth-grade math course and take an algebra course instead, he or she can administer a placement test that is aligned to the state standards for eighth-grade math. If the student demonstrates mastery on most of the eighth-grade content, then the student may be a good candidate for the algebra course. In some cases, however, it may still be the case that teachers responsible for the target course must identify the prerequisite content that should be included on the placement test.

ii. Psychometric Issues Regarding Placement Tests

In some contexts, a placement test will be administered to only one or two students at a given time. This means that the test must have strong psychometric properties that have been established in advance based on a much larger sample of students. For example, the placement test must be reliable. As discussed elsewhere in this document, reliability is a function of test length. ATI research indicates that tests of 70 items or more are likely to yield reliability coefficients in the .90s.Tests of 50 items can be expected to yield coefficients in the .80s. In order to maximize the accuracy of student ability estimates, the placement test should contain at least 50 items.

In addition to adequate test length, a placement test should be comprised entirely of items with well-established IRT item parameter estimates. The scale scores or DL scores that are generated by IRT analysis are based, in part, on the item parameter estimates that describe how each item functions, such as its relative difficulty. In most situations the item parameter estimates can be generated by the IRT analysis simultaneously with the student ability estimates represented by the DL scores. However, if a placement test is to be used for just a few students at a time, there will not be enough student performance data to generate item parameter estimates. Therefore, the item parameter estimates must be known before the assessment is used to generate student ability scores. Fortunately, the item parameter estimates can be derived in the context of other assessments. As long as the placement test is comprised of items with good, pre-existing item parameter estimates, it does not necessarily need to be piloted on large numbers of students before it can be used to generate accurate student ability estimates.

Precision of student ability estimates derived from a placement test may be maximized if the assessment is constructed to provide optimal information at a targeted range of ability levels. All assessments are subject to measurement error, and the degree of measurement error varies across the ability scale. For example, a pre-calculus assessment, no matter how well it is constructed, will not provide very accurate ability estimates for most third-grade students. Test information is the inverse of measurement error. Even within the target grade level, any assessment will provide greater information at some regions of ability than others. If an assessment such as a placement test is intended to focus decisions on a particular region of the ability scale, then the items on the assessment can be selected to maximize measurement precision at the target ability levels. For example, a test that is designed to decide whether students are to be placed in an advanced course may be designed to maximize measurement precision at or above one standard deviation above the mean. A placement test used to identify students for remediation may be designed so that measurement precision is maximized below one standard deviation below the mean. (Notice that when used in this manner a placement test is quite similar to a screening assessment.) A placement test that is designed to determine which of three possible grade levels is most appropriate for a new student should be designed so that measurement precision is maximized across as broad a range of ability levels. As long as the items that comprise the placement test have existing IRT item parameter estimates, a placement test can be designed to maximize measurement precision at the range of ability levels that is of greatest interest. CAT is an ideal solution for placement tests because it ensures that each student is assessed with an instrument that maximizes measurement precision at or near his or her ability level.

The decisions that are made based on student performance on placement tests typically depend on where the student scored relative to some cut score. The choice of where to place the cut score depends on the goals for classification. One approach is to administer the

placement test to all students in the grade, and then use percentile ranks to determine the top percent of students who are then placed in an advanced course. Although not optimal, this norm-referenced approach allows the district to align course enrollment with available resources. Generally speaking, however, a criterion-referenced approach is more appropriate for placement tests. In other words, students should be able to demonstrate a specific level of mastery of the content on the placement test before being recommended for an advanced course. If the content of the placement test is aligned with state standards and with the high-stakes statewide assessment, then cut scores on the placement test can be aligned to the high-stakes assessment as well and the performance-level classifications used by the state can be used to guide placement decisions. For example, a seventh-grade student at the end of the year may be given a placement test that is aligned to the eighth-grade math standards and to the eighth-grade math statewide assessment. The district may consider a score in the highest performance-level range on the assessment to be an indicator of eligibility to skip the eighth-grade math course and take an introductory algebra course instead. This approach requires that the placement test be administered to the full distribution of students in the grade level, at least initially, so that test equating procedures can be used to properly align the cut scores with the high-stakes statewide assessment cut scores.

Under some circumstances, the placement test may not be administered to the complete distribution of students. For example, a teacher or administrator may identify a subset of likely candidates for advanced placement based on course grades and/or performance on the high-stakes statewide assessment. A placement test would then be administered to this subset of students in order to gather an additional data point to inform the placement decision. In this situation, the cut scores must stand on their own regardless of whether the full distribution of student scores is present in the data set. Ideally, the placement test will have been administered to a complete distribution of students at some point in the past so that cut scores for the placement decision can be determined using test equating procedures. If this is not possible, it is common practice for the teachers who design the placement test, by identifying the prerequisite content for the target course, to also identify the cut score that represents adequate preparation for the advanced course. Regardless of the approach used to identify the cut scores for placement decisions, their appropriateness must be evaluated empirically.

Empirically evaluating validity is, of course, a concern for placement tests. However, the evaluation of the validity of placement tests is not always as straightforward as it is for many of the other assessment types discussed in this document. The assessment of validity is tied directly to an evaluation of the goals of the assessment. Many years ago, Messick (1989) put forth the view that validity is an evaluation of the appropriateness of decisions made based on student scores on the assessment. To properly evaluate the validity of an assessment, then, one must first clearly identify the intended use of the assessment. The intended use of placement tests within school districts is to place a student in a course or grade level in which he or she is likely to succeed. If a placement test is valid, then the students who are identified for placement in an advanced course should generally be successful in that course, and students who are identified as not being ready for the advanced course should be less successful in that course. It is immediately apparent that at least half of the data required for an evaluation of this type will never exist: if a student is identified as not being prepared for a course, he or she will most likely not enroll in it and so no evaluation can be made regarding relative success in the course.

Evaluation of validity of the placement test is further complicated by the question of how success in the course is best measured. It may be tempting to use the course grades assigned by the teacher as an indicator of student success in the course (e.g. Kane, 2006), but this is not

ideal because teachers may be inconsistent in how they assign grades. In many cases it may also be impractical to use student scores on the high-stakes statewide assessment for the evaluation of placement test validity because in most states[1] the statewide assessment that is taken by the student depends on the student's grade level, not the content that the student is actually studying. For example, the eighth-grade student who has been placed in an algebra class will, at the end of the year, take the eighth-grade, high-stakes statewide assessment that is aligned to the eighth-grade math standards, not the algebra content he or she has been studying. The student's performance on the statewide assessment will therefore provide no indication of whether the student was appropriately placed and achieved success in the algebra course.

The best option for evaluating the validity of the placement test as a tool for placement decisions may be to evaluate the student's performance on district-wide benchmark or interim assessments that are aligned to the algebra standards and which are taken by all students taking algebra, whether as their normal grade-level course or as an advanced placement. If the student is scoring at or above proficient on the benchmark assessments, then it can be concluded that the placement decision was appropriate and that, therefore, the placement test is valid with regard to its intended purpose. Further, if enough students are taking the course as an advanced placement, then the forecasting accuracy of the placement test can be assessed. The assumption would be that students who scored above the relevant cut score on the placement test are likely to score at or above proficient on the benchmark assessments for the targeted course. The proportion of students scoring at or above proficient on the benchmark assessment will provide an indication of whether the content and/or cut scores on the placement test need to be re-evaluated.

iii. Assessment Procedures for Placement Tests

The procedures for administering placement tests may vary with the context of the assessment. If the assessment is being administered to a large group of students, such as an entire grade level in a district or school, then the testing window should be as narrow as possible for two reasons. First, a narrow testing window enhances test security in that there is a shorter time frame during which teachers and students have access to the assessments. Secondly, a narrow testing window ensures that all students are being assessed at the same point in their educational development, so that placement decisions for all students are based on scores that reflect a common time frame for opportunities to learn during the school year. Otherwise, students who are assessed later in the year would be more likely to be placed in advanced courses than students assessed early on, at a time when they have been exposed to less instruction.

A placement test may also be an instrument that is kept on hand and administered to new students as they enroll in the school or district. Under these circumstances the time frame for the administration of the assessment is open ended. Placement decisions will be based on the degree of mastery of material at various grade levels rather than in direct comparison to the performance of peers. The time of year should be considered when such placement decisions are made. For example, consider a new student whose performance on the placement test indicates that he or she has mastered most of the third-grade material but very little of the fourth-grade material. If the assessment had been administered in August or September, the decision would probably be to place the student in the fourth grade. However, if the assessment had been administered in January, it may be more appropriate to place the student in the third

[1] California's CST assessments in math beginning with grade seven are one exception to this rule.

grade so that he or she can benefit from a full year of instruction at the fourth grade level in the following year.

As with most assessments, placement tests should be kept as secure as possible. However, in some cases the very nature of the placement test precludes some of the normal security procedures. For example, a placement test that is kept on hand for administration to new students as they enroll will necessarily be stored, either electronically or physically, for a much longer time period than other assessments. Precautions should be taken to make sure that access to the stored tests is limited.

XI. Observational and Rating-Scale Assessment

The term observational assessment has been used to describe a wide variety of evaluation activities. Included under this broad umbrella are non-structured notes that a teacher might make when a student is struggling with a concept as well as highly structured evaluation of a work sample based on detailed criteria or rubrics that spell out specific scores for performance at different levels (i.e., rating-scale assessment). Many have touted the advantages of documenting student performance on a real life task. Observational assessment, and in particular rating-scale assessment, is also commonly used to assess the proficiency of teachers and other educators. The results of the assessments may then be used as part of educator evaluation and instructional effectiveness initiatives.

A. Purpose and Goals of Observational and Rating-Scale Assessment

Observational assessment is often used to evaluate learning of students of all ages. In the case of young children, obtaining reliable assessment results can be challenging. Observational assessment can provide a means of evaluating a child's skills as he or she goes about his daily activities. In the case of older students, observational assessment provides a method for delivering formative assessment in order to check for understanding as a lesson is delivered. Observational assessment is also used to evaluate the efficacy of programs or teachers.

In the case of young children, observational assessment is popular because of the widely recognized difficulty in obtaining reliable results using what are thought of as traditional testing methods in which a child is asked to respond to series of questions either one-on-one or in a group. By contrast, observational assessment allows for the child to be evaluated while doing normal day-to-day tasks in a variety of different settings. National Association for the Education of Young Children (NAEYC) guidelines call for frequent observation and evaluation of children's development that is targeted and takes advantage of many sources of information and includes careful documentation of the observations as they are made.

With older children, observational assessment can be used as a means of conducting formative assessment. Formative assessments are evaluations that are intended to guide instructional decisions. The National Mathematics Advisory Panel calls for the use of formative assessment in their 2008 report. They recommend the regular use of standards-aligned items to evaluate student understanding of the concepts that they are being taught. They point out that these assessments can be structured to take less than 10 minutes to deliver.

Observational assessment can be used not only to determine whether a student has acquired a skill, but also to establish how a student approaches a problem to be solved.

Observations of problem solving can reveal the steps that students apply to solve a problem and mistakes they may make that interfere with problem solving.

Observational assessment and, in particular, rating-scale assessment may be used to conduct staff evaluations as well as to document student learning. For instance, as part of a teacher evaluation program, a school principal may observe teachers in the school providing instruction in the classroom setting. A scoring rubric may be devised for evaluating teacher skills. The rubric will define the kinds of skills required for effective teaching. In addition, it will provide concrete examples of the application of those skills.

B. Observational and Rating-Scale Assessment Characteristics

Scoring is a particularly important issue in observational and rating-scale assessment. Consideration of the question that will be addressed with the assessment data is critical to determining the scoring approach that will produce the best results. In the simplest case of a teacher asking her students a quick question to check understanding before proceeding ahead with the lesson, scoring is straightforward. Did the children answer the question correctly? Summed scores are often used to summarize informal observational assessments such as responses to teacher questions. Calculation of summed scores can be automated when students respond using response pads.

Observational and rating-scale assessment may also be used in those instances in which the goal of assessment is to determine an individual's position on a developmental scale. For example, observational assessment is frequently used to assess the progress of young children on a developmental continuum. IRT techniques may be used to place observed capabilities on a continuous scale. Summed scores do not provide a continuous scale because they are affected by variations in the difficulty of the tasks being assessed. IRT adjusts for differences in task difficulty. Because IRT allows capabilities to be placed on a developmental continuum, it may be used to answer questions about what skills a student is prepared to learn next.

The reliability and validity of an observational or rating-scale assessment instrument may be evaluated using similar techniques to those utilized for more traditional assessments. For instance, IRT may be used to determine internal consistency. Validity data may be obtained by conducting analyses such as correlating results with state test data. If an observational or rating-scale assessment instrument has been designed with clear criteria for making ratings, is of sufficient length, and is appropriately targeted for the skill level of the individuals with whom it is being used, then useful results may be expected.

C. Observational and Rating-Scale Assessment Procedures

Observational and rating-scale assessment involves special challenges that influence the procedures implemented in carrying out this form of assessment. The most salient of these is that observational and rating-scale assessment invariably involves a degree of subjective judgment. The subjective nature of observation often becomes apparent when two people observing an event have strikingly different accounts of what actually happened. The time-honored procedure for addressing the problem of subjectivity is to have two or more people conduct the required observations. When two people agree as to what they have seen, their account enjoys a degree of objectivity that is not possible when there is only one observer. When possible, observational assessment should include provisions for multiple observers. Understandably, when many observations are made, observers may not agree all of the time.

Statistical procedures are available to measure the degree of agreement between observers. ATI typically uses latent-class analysis to measure agreement. The latent-class approach gives the probability of agreement among multiple observers. The involvement of multiple observers complicates the scoring process because there are at a minimum two sets of scores. In some cases, scores are taken from an observer designated as the primary observer. In other instances, the contributions of each observer are included in the score.

In many circumstances, it is not practical to require multiple observers for observational and rating-scale assessment. When this is the case, it is useful to find other ways to support the credibility of the observations. It is generally useful to document the source of the judgment as part of the effort to establish the credibility of the data. ATI provides users with source documentation tools as a component of its observational and rating-scale assessment technology. A variety of methods may be used to obtain evidence of proficiency including direct observation, the examination of documents such as lesson plans, information obtained during an interview, and information gathered from automated analyses of computer use.

A second approach that may be used to support the credibility of observational data is to establish the relationship between observational scores and other assessment results. For example, child observational assessment results may be correlated with results from a direct assessment. Likewise observations of teaching can be correlated to scores on student assessments.

Scheduling observational assessments often poses unique challenges that must be addressed to ensure the credibility of observational assessment results. The behaviors to be observed often occur over an extended time span because it may not be possible to schedule the occurrence of the behaviors directly. For example, observing a child's language skills typically requires observations occurring over a relatively long time period. When observation requires information on behaviors that occur naturally within a particular environmental context, observations are typically scheduled to occur over many months. For example, the development of young children's cognitive skills is often observed continuously by teachers over an entire year. ATI provides tools that accommodate continuous observation, observation occurring within designated periods, and observation at a particular point in time.

XII. Conclusion

A comprehensive assessment system serving the needs of students from preschool to the 12th grade must invariably include many different types of assessment. The various assessment types serve different purposes. Benchmark and formative assessments are designed to inform instruction. Pretests and posttests are useful for measuring academic progress. Analyses of instructional effectiveness assessments can identify classes and schools that are highly successful as well as those needing additional assistance. Screening instruments are useful in identifying students at risk for learning problems. Placement tests inform grade-level placements, and advanced course placements. Computerized adaptive tests provide efficient measures of academic proficiency. Automated construction of Computerized adaptive tests increases testing options and enhances the capability to meet unique local needs. Observational and rating-scale assessments provide authentic measures of competencies in the environment in which those competencies are used. In addition, they provide immediate information that can be used to guide instruction.

As the extensive list in the preceding paragraph implies, the most obvious benefit of a comprehensive assessment system is that it helps to ensure that all of the school's assessment needs are met. Insofar as possible this is accomplished by including all of the various required types of assessment within the system. The vast majority of assessment types will be native to the system. However, there may be some types that are imported into the system. Statewide test results provide a familiar example.

An additional benefit to a comprehensive assessment system is that it creates economies in the assessment process including the construction, publication, administration, scoring, and reporting technology necessary for system implementation. This is the case because the same technology can be used for many different types of assessments.

Perhaps the most important benefit of a comprehensive assessment system is that it supports the ability to adapt to continuous change, which is a hallmark of education in our times. Assessment needs change continuously. A well designed comprehensive assessment system includes technology capable of accommodating change. For example, a well designed system should include the ability accommodate continuously changing standards. It should have the capability to rapidly align items to those standards. It should include dynamic item banks that expand continuously. It should have the capability to generate innovative item types that will be required as the transition to online assessment accelerates, and it should be capable of incorporating new types of assessments to meet changing needs. A comprehensive assessment system should always be and will always be a work in progress. Nothing less will meet the educational challenges faced by educators in today's rapidly changing world.

XIII. References

Bergan, J.R., Bergan, J.R., & Burnham, C.G. (2009). *Benchmark Assessment in Standards-Based Education*. Assessment Technology, Inc. Available for download at http://www.ati-online.com/galileoK12/K12Research.html.

Bergan, J.R., Bergan, J.R., Burnham, C.G., Callahan, S.M., & Feld, J.K. (2011). *Instructional Effectiveness Assessment*. Manuscript in preparation. Draft available by request by contacting Assessment Technology, Inc. at http://www.ati-online.com/forms/contactus.asp.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal*, 28(3), 683–714.

Black, P. and Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: principles, policy and practice, 5, 7-74.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Compton, D.L., Fuchs, D., Fuchs, L.S., & Bryant, J.D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394-409.

Edwards, M.C. & Thissen, D. (2007). Exploring potential designs for multi-form structure computerized adaptive tests. In D.J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement* 4th ed., pp. 17-64. Westport, CT: American Council on Education and Praeger.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement*. (3rd ed., pp. 13-103.). New York: American Council on Education and Macmillan.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system*. Aspen, CO: Aspen Institute.

Raudenbush, S. W. Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. 2nd edition. Newbury Park, CA: Sage.

Scriven, Michael. (1967).The methodology of evaluation. In Gredler, M. E. *Program Evaluation*. (p. 16) New Jersey: Prentice Hall, 1996.

U.S. Department of Education, National Center for Education Statistics. NAEP Validity Studies: Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP, NCES 2003–14, by R. Darrell Bock and Michele F. Zimowski. Project Officer: Patricia Dabbs. Washington, DC: 2003

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2005). Statistical models for categorical data. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 1-14). Mahwah, NJ: Erlbaum.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Webb, N. (2006). Identifying Content for Student Achievement Tests. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 155-180). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Whitney, D.R. (1989). Educational Admissions and Placement. In R.L. Linn (Ed.), *Educational Measurement*. (3rd ed., pp. 515-526.). New York: American Council on Education and Macmillan.