



The Research Base for PLATO[®] Test Packs with Prescriptions on the PLATO Learning Environment[™] (PLE[™])

Technical Paper #19

Kim Baker

Senior Curriculum Specialist

PLATO Learning
10801 Nesbitt Avenue South
Bloomington, MN 55437

800.44.PLATO
www.plato.com

November 2007

Contents

Executive Summary	3
Assessment	3
High-Stakes Tests	4
Low-Stakes Tests	4
Criterion-Referenced Tests.....	4
Norm-Referenced Tests.....	4
Trends in Assessment	5
<i>Testing for Accountability</i>	5
<i>Testing to Make Educational Decisions</i>	5
Evaluate Student Progress	6
Evaluate Student Mastery.....	6
Evaluate Instructional Programs	6
Challenges in Assessment	6
<i>Reporting Inefficiencies</i>	6
Delayed Results.....	7
Inadequate Information.....	7
<i>Standards and Multiple Measures</i>	7
State Standards	7
Alignment of Assessments.....	8
<i>Test Limitations</i>	8
Test Quality	8
Clear Purpose.....	9
The PLATO® Test Packs with Prescriptions Solution	9
<i>Efficient Reporting</i>	10
Authentic Assessments.....	10
Flexible Implementation.....	10
<i>Informative Assessments</i>	11
<i>Best Practices for High-Quality Tests</i>	13
Best-Practice Methods in Assessment.....	13
Readability.....	14
Comparability	14
Key Research-Based Principles Driving PLATO Test Packs	14
State-Specific Fixed Benchmark Tests.....	15
National Cumulative Fixed Benchmark Tests.....	15
National Progress Series Tests.....	15
Reports.....	16
Prescriptions.....	17
Conclusion	17
Appendix A: PLATO Learning Item Writing Guidelines	18
Appendix B: Readability Measures	25
Appendix C: Comparability of Test Versions	28
References	31

Executive Summary

This paper discusses the research basis for PLATO® Test Packs with Prescriptions on PLE™. It describes two trends in assessment: (1) testing for accountability, and (2) testing to make educational decisions. This paper also conveys the research around several key challenges in assessment:

- Delayed test results
- Ineffective reporting
- Variations in academic standards quality
- Issues with alignment of assessments to each other
- Overall test limitations

In addition, this paper presents the ways that PLATO Test Packs addresses these challenges. They include the following:

- Offering assessments that use time efficiently and allow for flexible implementation
- Developing items and tests according to respected professional academic organizations, national standard setting bodies, and standards evaluation organizations
- Following the rigorous standards and guidelines set forth by leading assessment experts and the Standards for Educational and Psychological Testing to create high-quality tests
- Developing a variety of tests to address various needs such as unit and cumulative test data
- Providing tests within a reporting environment that allows educators to efficiently use student data
- Offering prescriptions for remediation in areas in which students need support

PLATO Test Packs fills a specific need: to help administrators by providing classroom teachers with a tool to measure student achievement based on state or national standards at a given period of time. PLATO Test Packs provides a series of criterion-referenced, low-stakes tests, which may be used in either a formative or summative manner.

Assessment

Assessment is the comprehensive term for a range of processes to gain information about student learning (Linn and Miller 2005). A test is a particular type of assessment given at a fixed point in time. The purpose of assessment is to ultimately improve student learning. Assessment procedures can be classified in terms of their role in the classroom. There are two key categories among several types of assessments:

1. **Formative assessments** monitor progress during instruction. They help teachers make decisions about student learning and provide ongoing feedback to students. The results of formative assessments can be used as a measure of how well students are progressing toward a goal.

2. **Summative assessments** assess achievement at the end of instruction. They measure performance against the intended learning goals after they are taught. The results of summative assessments can also be used to make long-term curriculum and professional development decisions.

Most tests can be both formative and summative, depending on how they are used. For example, tests can be implemented to provide feedback to students and teachers during learning or to measure progress toward learning outcomes after instruction. For accountability purposes, assessments are almost always summative. Educators tend to classify tests as either high-stakes or low-stakes.

High-Stakes Tests

All tests are used to make decisions. If the decision has major consequences—such as determining graduation, promotion, or employment—the stakes are high. High-stakes tests are often used to hold schools and districts accountable using a summative measure. High-stakes tests are typically externally mandated. Most often, state standards tests fall into this category. Other achievement tests such as the Stanford Achievement Test and the Iowa Test of Basic Skills are both high-stakes tests. High-stakes tests are held to the highest quality expectations because the consequences of an incorrect decision are large and lasting.

Low-Stakes Tests

In contrast, if a decision has minor consequences, such as testing for a grade or placement within a curriculum, then the stakes are low. Low-stakes tests are often used to find a measure of student performance against some learning objective or intended learning goal.

It is important to recognize that formative tests are not always low stakes tests and summative tests are not always high stakes tests. In addition, providing a particular assessment is just the first step in improving learning. For example, one essential step in using formative assessments is determining what to do with the data.

High-stakes and low-stakes tests can be put into two categories: (1) criterion-referenced test and (2) norm-referenced tests.

Criterion-Referenced Tests

Criterion-referenced tests are scored by comparing performance to pre-established criteria, such as objectives, benchmarks, or learning goals. The purpose of criterion-referenced tests is to determine how well students have learned specific information. Most states use their state standards tests, which are intended to be aligned with their state's academic standards. For example, California offers the California Standards Tests (CSTs). The California academic standards are the basis for these tests.

Norm-Referenced Tests

In norm-referenced tests, scores are ranked and compared to each other. The results of these tests compare a student's score to scores from a sample of representative students.

Norm-referenced tests are not designed to measure student knowledge of a specific curriculum. Rather, they are designed to classify students according to what they know compared to other students. Norm-referenced tests are often used to measure basic skills that students are expected to have. The SAT test and the Stanford Achievement Tests are examples of norm-referenced tests.

Trends in Assessment

For more than a decade, educators have worked to set challenging academic standards and to use assessments to measure student progress toward meeting those standards. For this reason, educational assessment is playing a greater role in making decisions than ever before (National Research Council 2001). The central purpose of assessment is to answer the question, “How well does the student perform?” (Linn and Miller 2005). The information is then used to improve student learning. Within that goal, there are two secondary purposes: (1) testing for accountability and (2) testing to make educational decisions.

Testing for Accountability

Prior to the No Child Left Behind (NCLB) Act, many states required that all students take summative assessments to measure performance on basic skills. The NCLB Act requires that states test all students in grades 3–8 and in at least one grade in high school in reading or language arts and mathematics every year (Linn and Miller 2005). As of the publication date of this document, beginning in the 2007–08 school year, states are required to test students annually in science in at least one grade within three grade spans (grades 3–5, grades 6–9, and grades 10–12). These high-stakes tests results will be used to hold districts and schools accountable.

One way districts and schools are held accountable is through adequate yearly progress (AYP) targets. NCLB legislation provides federal funds to states to help educate students. In order to determine if districts are on track to meet the goal of student proficiency in various subjects, each state sets benchmark goals to measure adequate yearly progress towards proficiency. AYP targets attempt to define how students are progressing. When schools do not meet defined AYP benchmarks, punitive sanctions are placed on the school, and the school is given a set timeline to improve student learning before losing federal funds.

Nearly all schools and districts are adhering to NCLB requirements and administering high-stakes achievement tests. These tests provide a summative, criterion-referenced (state standards tests) or a norm-referenced (Stanford tests; Harcourt Assessment) measure of student performance. It is evident that testing for accountability purposes has become an integral part of educational reform.

Testing to Make Educational Decisions

In addition to using high-stakes tests to hold districts accountable, both high-stakes and low-stakes tests have long been employed to measure student performance in order to make short-term and long-term educational decisions. There are three ways testing can be used to make decisions.

Evaluate Student Progress

First, tests are used to evaluate student progress. When assessments are used in a formative manner, teachers check progress against intended learning objectives during instruction. Educators can then make decisions about areas where students need intervention, re-teaching, or alternative styles of instruction. These decisions can help inform and guide instruction to meet the needs of all students, wherever they are in their learning process.

Evaluate Student Mastery

Next, tests are used to evaluate student mastery. At the end of a unit, lesson, or class, teachers measure students in a summative manner. Evaluating student mastery gives teachers a view of student achievement against the intended learning goals. This summative view provides a clear picture of how well students learned what was taught.

Evaluate Instructional Programs

Finally, tests are used to evaluate instructional programs. Schools can determine if their instructional programs are working by analyzing test data in terms of where student performance is against where the district would like it to be (Bernhardt 2003). By identifying these gaps and the root causes of these gaps, schools can develop improvement plans specifically aimed at areas of weakness. In addition, administrators can plan for teacher professional development needs based on student performance gaps.

High-stakes test results provide one measure of student achievement at a given point in time. Other student assessment procedures, such as low-stakes tests, quizzes, performance assessments, etc., help schools gain a comprehensive picture of student achievement. By using both summative and formative test results, schools can make curriculum, professional development, and resource decisions.

Challenges in Assessment

The increased emphasis on assessment means that several key challenges in assessment are more prevalent than ever:

- Reporting inefficiencies
- Deficient assessments
- Test limitations

Reporting Inefficiencies

Given that the purpose of assessment is ultimately to improve learning, the timeliness of test results becomes essential in order to provide feedback to students and teachers in an effective, efficient way (National Research Council, 2001). There are two challenges within this feedback: (1) test results are delayed, and (2) when reports are delivered, they offer too little information.

Delayed Results

Due to NCLB pressures, schools must find more efficient ways to test students and learn the results of those tests quickly (Sharkey and Murnane 2003). For test data to be effective in guiding instructional decisions, testing must be non-intrusive on class instruction, and the results must be clear and immediate. One challenge with high-stakes tests is that their impact is delayed because results are often not available until the summer following the administration of the test. The delay in test results causes a delay in feedback and, thus, a delay in remediation for students.

Inadequate Information

In addition, high-stakes state test results often provide too little information to help improve student achievement (Rettig et al. 2003). The student data is aggregated. With the results in this format, it takes teachers and administrators a long time to wade through the reports. This process is inefficient. In addition, some of the data gives educators general information that they already know about student learning. For example, teachers most likely already know who the students are who are struggling in a particular subject. Thus, these high-stakes assessments by themselves have little impact on individual students. Educational leaders and teachers need accurate, detailed, and regular information about which students are mastering intended learning outcomes in the curriculum (Jerald 2003). It is often these tests (formative assessments) that are most suited to guide improvements in student learning (Guskey 2003).

Within formative assessments, one challenge is the process of acting on reported data. One key question that arises following test reports is often, “What do I do with this data?” Educators grapple with this question and often plan trainings and professional development around it. While this challenge isn’t specifically addressed in this paper, it is important to recognize this challenge when discussing reporting inefficiencies.

Standards and Multiple Measures

Unsatisfactory state standards result in low-quality, unsatisfactory state assessments. In addition, some state assessments are out of alignment with local and national assessments.

State Standards

Standards-based reform is by far the most significant movement in American K–12 education today. Good standards matter more now than ever before because the NCLB Act has placed great emphasis on meeting standards. Schools whose high-stakes test results do not show progress toward attaining these goals face serious consequences (Finn, Julian, and Petrilli 2006).

However, the quality of state standards and the tests that are derived from them have received criticism from professional agencies. For example, state standards not only vary widely from state to state, but they are, on the whole, mediocre according to the Thomas B. Fordham Foundation. Although states are making efforts toward continuous improvement of their standards, two-thirds of the nation’s K–12 students attend schools in states with C-, D-, or F- rated state standards according to criteria used by the Thomas B. Fordham Foundation (Finn, Julian, and Petrilli 2006; Gross et al. 2005; Klein et al. 2005; Mead, Finn, and Davis 2006; Stotsky and Finn 2005).

State standards play an important role in what happens in classrooms. For example, many districts create curriculum guides that are similar or identical to their state standards. Administrators and teachers work together to determine pacing and timing for teaching these standards. The high-stakes, state standards tests that schools are held to are based on the academic learning standards for their state. In this way, each state's standards serve as a set of intended learning outcomes that guide instruction and that are used to measure student progress through high-stakes assessments.

Using multiple measures to determine student achievement is valuable and necessary in painting a comprehensive picture of student performance. However, when multiple measures are used for the same purpose, without consideration of the varied purposes of each measure, the picture of student performance is less than comprehensive.

Alignment of Assessments

One key challenge is that many assessments are not aligned with other assessments. For example, high-stakes assessments are often out of line with low-stakes and formative assessments (National Research Council 2001). This lack of alignment results in a less than comprehensive picture of student achievement. It also points to challenges teachers face in ensuring that their students are progressing toward mastering state standards. For instance, it is possible for students to achieve particular outcomes on high-stakes tests and quite different outcomes on low-stakes tests and vice versa.

A comparison of national tests and state tests shows that they are also out of alignment. For example, the National Assessment of Educational Progress (NAEP) national test has been charged with determining national student achievement in several subjects for students in grades 4, 8, and 12. The test differs greatly from state tests in several ways. One key difference is the available item types. NAEP offers open-ended item types, including essays and drawings. State tests are gradually incorporating more varied item types, but the majority of high-stakes state tests use multiple choice items. As a result, there is a divide between the types of skills and the difficulty levels of skills that can be measured on the NAEP national test and on state tests. The NAEP tests, for example, have item types which are appropriate for measuring higher order skills, relationships, and creative problem solving. Because the majority of state tests use primarily multiple choice items, the objectives that can be best measured are often more fact based and do not span the taxonomy of cognitive knowledge.

Test Limitations

Tests are limited. Recognizing the limitations of tests is a critical component to the proper use of results (Linn and Miller 2005).

Test Quality

Assessment experts describe test creation as an art. Test development takes great skill. It requires a good grasp of subject matter, a clear understanding of desired learning outcomes, a psychological understanding of students, sound judgment, persistence, and creativity (Linn and Miller 2005).

Research shows that a high-quality test is one that is the clearest measure of intended learning outcomes. This means that each item on the test is the clearest measure of the intended learning outcome.

Creating test items that meet this criteria is a challenge because there are many potential barriers:

- Ambiguous statements
- Excessive words
- Difficult vocabulary
- Unclear instructions

In addition, high-quality tests are free of bias. When assessments contain bias, the likelihood that the item or test is the most direct measure of the intended learning outcome goes down.

Clear Purpose

The purpose of assessments is to gain information about student learning in order to make instructional decisions to improve learning. For this reason, assessments are considered a means to an end rather than an end themselves (Linn and Miller 2005).

Even when employing a test of the highest quality, it's important to recognize that the test has limitations. No one type of assessment fits all assessment purposes. And a test itself is only one element in the assessment process. The best information that can be obtained on student learning is information that comes from a combination of sources, including classroom assignments, formative assessments, and high-stakes assessments (Gandal and McGiffert 2003).

Experts describe good assessment as an ongoing process of reasoning from evidence (National Research Council 2001). Generally, the more sources of accurate, appropriate evidence available, the more likely the reasoning is sound.

The PLATO[®] Test Packs with Prescriptions Solution

PLATO Learning's test design process addresses test limitations in several ways. First, high-quality tests are composed of high-quality items. PLATO Learning developed rigorous guidelines for creating high-quality items. These guidelines, which were derived from leading assessment authorities Thomas M. Haladyna, Robert L. Linn, and M. David Miller (see [Appendix A](#)), were documented in a robust item writing guide. In addition, many trainings were conducted to ensure that developers understood and applied these guidelines. PLATO Learning also created acceptance criteria (see [Appendix A](#)), which is comprised of minimum quality expectations in the form of a rubric. Instructional designers and curriculum specialists reviewed items using the rubric. For more information on PLATO Learning's item writing guidelines, please refer to the [Best-Practice Methods in Assessment](#) section of this paper for a summary or to [Appendix A](#) for more specific information. Finally, the PLATO Learning item and test development process

includes rigorous peer reviews, subject matter expert reviews, and editorial reviews to reduce bias and clarify the purpose.

PLATO Test Packs provides administrators and teachers with achievement data by providing student achievement data based on state or national standards. With this data, teachers can make decisions about interventions in instruction, can make modifications to the instructional approaches taken, and can consider alternative methods to help students meet learning goals. In addition, district leaders can use this data to make curricula and professional development decisions.

Efficient Reporting

Challenge: Inefficiencies in Reporting

Given that the purpose of assessment is ultimately to improve learning, the timeliness of test results becomes essential in order to provide feedback to students and teachers in an effective, efficient way (National Research Council 2001). To raise student achievement, schools need test results in a timely manner (Gandal and McGiffert 2003).

The PLATO Test Packs Solution: Efficient Reports

Using technology-driven assessments has several advantages. First, scores can be obtained in a timely manner. One of the key disadvantages of high-stakes, paper-based tests is that results are often not available until students have finished the school year. This delay means that, in most cases, schools cannot make instructional changes until the following school year. Using PLATO Test Packs eliminates the time delay associated with administering and scoring paper-and-pencil tests. PLATO Test Packs provides automated online scoring, which allows educators to make immediate decisions that will impact learning. In addition, PLATO Test Packs tests are developed using state and national objectives. Reports show areas where students need intervention in order to meet objectives.

Authentic Assessments

High-stakes tests by themselves often provide too little information to improve student achievement (Rettig et al. 2003). For example, data is often reported at a subject or strand level, rather than at an objective level. The best information on a student's progress comes from a combination of low-stakes, high-stakes, and classroom assessments. The ease of administering and scoring PLATO Test Packs tests saves time and allows teachers to focus on essential instructional areas and other types of authentic assessments.

Flexible Implementation

PLATO Test Packs offers the flexibility for teachers to choose between using tests in a formative or summative manner. If tests are used in a formative manner, instructional intervention can be provided at the point when it is needed. Similarly, if PLATO Test Packs tests are implemented in a summative manner, educators can use the data to gauge student performance prior to the high-stakes test. No matter which approach is taken, the appropriately designed assessments can help teachers by revealing areas where students, teaching practices, and schools need to improve.

In addition, PLATO Test Packs is coupled with a reporting environment that is focused on informing instruction and demonstrating student progress. District leaders can also trace trends in data over time in order to reveal areas where curricula changes are needed.

Informative Assessments

Assessments provide students with information about what is expected or important. Useful assessments must make clear what is being measured, and they must measure what educators value most (Gandal and McGiffert 2003). PLATO Tests Packs address two key issues within the realm of informative assessments: state standards and alignment of assessments.

Challenge: Varied State Standards

State standards vary in quality, yet they strongly influence what happens in classrooms. Good standards matter more now than ever before because the NCLB Act has placed great emphasis on meeting standards. Schools whose high-stakes test results do not show progress toward attaining these goals face serious consequences (Finn, Julian, and Petrilli 2006).

The PLATO Test Packs Solution: Up-to-Date State Standards

The criteria used to develop PLATO Test Packs tests are similar to state and/or national standards, which provide teachers with objectives that are likely to appear on high-stakes tests. PLATO Learning monitors these standards on an ongoing basis to ensure that educators have the most up-to-date assessments. State standards often undergo continual improvements and changes. PLATO Learning updates these standards within the management system once changes are approved by the department of education.

In addition, the criteria used to develop PLATO Test Packs tests are drawn from respected professional organizations, such as state departments of education and these national standard-setting bodies:

- National Council of Teachers of Mathematics (NCTM)
- National Council of Teachers of English (NCTE)
- National Academy of Sciences (NAS)
- National Council for the Social Studies (NCSS)
- National Center for History in the Schools (NCHS)

PLATO Test Packs tests are also based on information from these standards evaluation organizations:

- The Thomas Fordham Foundation
- The National Assessment Governing Board (NAGB)

In addition, PLATO Learning follows a rigorous development procedure based on the Standards for Educational and Psychological Testing, as set forth by the following organizations:

- American Educational Research Association (AERA)
- American Psychological Association (APA)
- National Council on Measurement in Education (NCME)

A rigorous development criteria creation process (see *Appendix A*) enables PLATO Learning to provide a solid foundation of assessment content.

One central shift in the area of assessment has been reducing the “gotcha” factor. In the past, assessments were viewed as a secret. Thus, the central thought running through students’ minds was, What is going to be on the test? Fortunately, educators and publishers have moved toward a more sensible approach in assessments—make the goals and objectives clear and measure them fairly. This informs educators of the full domain of items that appear on the test. Assessments that are meaningful do not surprise students. Instead, they reflect the concepts that were emphasized during instruction (Guskey 2003).

PLATO Test Packs tests are informative assessments in that the categories of learning goals for which the assessments are built are clear and available to the teacher prior to students taking the test. These categories are available within the details view of the test, which allows for a fairer test with fewer secrets or “gotchas.”

Challenge: Assessments Are Not Aligned

Many high-stakes tests are not aligned with NAEP or with low-stakes tests (National Research Council 2001). This lack of alignment results in a less than comprehensive picture of student achievement.

The PLATO Test Packs Solution: Aligned Assessments

One of the key principles of good assessment is that assessments must be clearly aligned with learning goals (National Research Council 2000). PLATO Test Packs offers tests that are clearly aligned to learning objectives and provide easy and immediate feedback for the student and teacher. Use of PLATO Test Packs also gives the teacher time to focus on other important areas of classroom assessment, such as performance assessments and observational assessments. For example, the Strengths and Needs Report groups students that struggle with the same standards. This saves the teacher time from manually analyzing and highlighting these groupings.

In addition to providing educators with assessment items that are based on state and national standards, PLATO Test Packs provides low-stakes assessments. These low-stakes assessments help provide a comprehensive picture of student learning.

Finally, PLATO Test Packs tests are similar in grade level and readability to other frequently used tests. This similarity is addressed in more detail in the Best-Practice Methods in Assessment section.

Best Practices for High-Quality Tests

Challenge: Test Limitations

Research shows that a high-quality test is one that is the clearest measure of intended learning outcomes. Test creation is complex because it requires a good grasp of subject matter, a clear understanding of desired learning outcomes, a psychological understanding of students, sound judgment, persistence, and creativity (Linn and Miller 2005).

The PLATO Test Packs Solution: High-Quality Tests

In addition to monitoring and updating standards on an ongoing basis and providing the efficient, accurate feedback that is needed to help educators make instructional decisions, PLATO Test Packs was developed using procedures and design standards that follow accepted professional practices for low-stakes, criterion-referenced tests. Development processes address best-practice methods in assessment writing, readability, and comparability.

Best-Practice Methods in Assessment

PLATO Test Packs tests were developed using best-practice criteria emphasized by leading assessment expert Thomas M. Haladyna (Haladyna, 2004). Additional assessment best-practice guidelines were derived from *Measurement and Assessment in Teaching* (Linn and Miller, 2005).

The central goal in test writing is to write a series of high-quality items. High-quality items closely measure the intended learning objective and lack ambiguity. There are several barriers to meeting this objective, including ambiguous statements, excessive words, difficult vocabulary, unclear instructions, and bias (Haladyna, 2004).

PLATO Learning has followed principles set forth by leaders in assessment to develop guidelines that help reduce ambiguity and ensure that each item matches the intended learning goal. The first guideline is to select the most appropriate item type (multiple-choice, true/false, matching, or grid response) to meet the stated learning objective. See [Appendix A](#) for more details about this guideline.

Developers have established clear guidelines for each item type. Several standard procedures ensure that these guidelines are followed for all new items:

- First, robust writing guides were created based on item-writing principles. These guides include an item-writing guide as well as subject-specific style guides for each content area: math, English/language arts, science, and social studies. See [Appendix A](#) for examples of PLATO Learning's guidelines on item writing.
- Second, targeted trainings around these guidelines were conducted for all new item writers and reviewers
- Next, a quality measure, or acceptance criteria, was created that defines the minimum quality accepted for an item. If an item does not meet the criteria, PLATO Learning does not use it on a test.
- Adherence to PLATO Learning's guidelines are tracked in order to determine needs for further training.

Finally, each subject area has a designated subject matter expert (SME), who reviews the items, organizes necessary resources, and plans trainings based on item reviews to further ensure quality.

Readability

All passages are evaluated and given readability levels. PLATO Learning verifies the suitability of language arts/reading levels for passages through automated and manual methods. For more information on readability, see [Appendix B](#).

Comparability

PLATO Test Packs tests were developed using a set of rules to create comparable items, tests, and passages. See [Appendix C](#) for these rules. Teachers can check student progress against standards or objectives using comparable test forms. The different test versions are comparable tests. That is, they test the exact same objectives and include a similar number of items per objective. The items also have comparable levels of difficulty between the two tests.

Through these processes to ensure best-practice methods of test creation, PLATO Test Packs provides high-quality, low-stakes tests.

Key Research-Based Principles Driving PLATO Test Packs

Principle 1:

Assessment is viewed as the primary tool for educational reform.

Historically, educational testing has been pervasive in U.S. schools with national tests, ITBS, Stanford 9, and Terra Nova. Due to the testing required by the NCLB Act, increased emphasis has been placed on assessment results in order to hold schools, districts, and teachers accountable. Furthermore, schools want to have data on student performance before students take high-stakes tests in order to help students succeed. For example, if a teacher has information showing that students in her class are struggling with an objective, the teacher can use this information to address this concept prior to students taking the high-stakes test.

There are several reasons why assessments are often used as the gauge for high-quality education. First, tests and assessments are fairly inexpensive when compared to other factors that impact student performance, such as smaller class size or professional development. Moreover, as policies or standards change, tests and assessments can be updated in a timely manner. Finally, our culture is familiar with observing results in terms of numbers. Test scores are visual and are typically reported by media and educational institutions (Linn and Miller 2005).

How Principle 1 is embodied in PLATO Test Packs

PLATO Test Packs offers a variety of test types to help meet district needs. Each test type within PLATO Test Packs serves to provide different information about student learning, which may be used for different purposes.

State-Specific Fixed Benchmark Tests

State-specific fixed benchmark tests are designed as a series of multiple, comparable tests per grade. Fixed benchmark tests are developed using state standards for that state. Because they are comparable at the objective level, these tests can be used at multiple points throughout the year to get a formative picture of how students are doing in satisfying learning objectives. Using this data, teachers can make decisions for a class or for individual students on how to intervene before high-stakes assessments take place. Each test in the series is meant to take approximately one class period to administer, which keeps the impact on instructional time manageable while still providing a reasonable alignment against the state standards.

National Cumulative Fixed Benchmark Tests

National cumulative fixed benchmark tests are designed as a series of two to three comparable tests per grade in five subjects—reading, writing, math, science, and social studies. The cumulative tests were developed against a consolidation of information from these sources:

- National standards
- Current Fordham Foundation A-rated state standards¹
- National standard setting bodies, including the National Council of Teachers of Mathematics (NCTM), National Council of Teachers of English (NCTE), National Academy of Sciences (NAS) National Council for the Social Studies (NCSS) National Center for History in the Schools (NCHS)

The State-Specific and National Cumulative Fixed Benchmark tests (FBT) can be used in a summative manner in that they are intended to determine the extent to which learning goals have been achieved. While summative assessments are often used to assign a grade or score, they also provide information to determine the appropriateness of learning objectives and the effectiveness of instruction (Linn and Miller 2005). PLATO Test Packs State-Specific and National Cumulative FBTs can also be used as formative assessments. They can be implemented to monitor progress during instruction, which allows teachers to make decisions about student needs.

National Progress Series Tests

Progress series tests provide assessments that are structured and sequenced in a way that is similar to middle school and high school courses with the focus on objectives within an instructional unit of the course. Each series of two comparable tests assesses understanding of a portion of the course objectives. The objectives are clustered together

according to the curriculum framework, similar to an instructional unit lasting four to eight weeks. Rather than testing the content of the whole course cumulatively, the progress series tests allow for more frequent testing just before or immediately after teaching the content of a unit in a course or textbook.

Principle 2:

The role of assessment is to provide a better education for the learner.

The central purpose of assessment is to improve student learning by providing classroom teachers with a measure of student performance during a given period of time. The idea is that when given meaningful data, teachers can respond to individual student educational needs (Gandal and McGiffert 2003).

To paint a comprehensive picture of student performance, teachers compile information from various sources: informal observations, classroom assessments, and high-stakes assessments (Gandal and McGiffert 2003). Assessments that teachers administer on a regular basis are the assessments that are most likely to increase improvements in student learning (Guskey 2003). One of the reasons for this likelihood is that teachers know classroom assessments are tied directly to instructional goals on scope and sequence. In addition, teachers gather regular information about which skills students are mastering, and this only comes with regular assessment.

Good assessments do not necessarily require complex assessment procedures. Multiple-choice tests, for example, can be used in ways that effectively assess student understanding (National Research Council 2000). What is needed is continuous feedback to students *and* teachers.

How Principle 2 is embodied in PLATO Test Packs

Reports

The ease and efficiency with which PLATO Test Packs can be delivered and scored provides teachers with an ongoing formative or summative view of student performance. In addition, PLATO Test Packs features a reporting environment that allows administrators and teachers to efficiently view student performance. PLATO Learning's demographic summary report allows educators to view demographic data on students. Educators can track data across schools within a school district and monitor progress within a class against selected demographic elements. The variety and flexibility of views helps provide the necessary information on which to make curricular and professional development decisions.

Furthermore, progress reports provide teachers with a view of their class or individual students in order to capture progress. Teachers can use this information to form a complete picture of student performance and make intervention decisions. PLATO Test Packs efficiently allows teachers to view student performance, to draw conclusions about areas of instructional intervention, and to make instructional decisions. Teachers can use this information to determine how to best differentiate instruction and to implement alternative modes of learning.

Prescriptions

In addition to an efficient reporting environment, PLATO Test Packs focuses on individual student performance by producing individualized remediation plans, or prescriptions. Prescriptions provide students with instructional content for items on which they did not perform well. For many of the fixed benchmark tests, PLATO Test Packs features a prescription that matches student performance needs with an appropriate instructional path. Students are given feedback when they most need it. In addition, teachers are given clear student assignments that require no additional customization, though the system allows for customization.

Along with reports that demonstrate student achievement, prescriptions give teachers valuable information about student performance to help modify instructional practices.

Conclusion

PLATO Test Packs was developed to address a specific need: to provide classroom teachers with information about student progress toward a learning goal. Three key challenges in assessment are addressed within the PLATO Test Packs solution: reporting inefficiencies, deficient assessments, and test limitations.

PLATO Test Packs addresses these three problem areas through a research-based design that includes best practices in assessment and takes advantage of technology to provide efficient, high-quality tests. Using this solution, educators can do the following:

- Obtain student data on performance against learning objectives in a timely manner
- Make informed instructional decisions using student data
- Use their valuable time on other measures of student progress
- Use data to make curricula and professional development decisions
- Implement remediation and interventions for students prior to high-stakes test

Using the best available research, the development team for PLATO Test Packs uses the best of technology to provide a resource that can dramatically inform, guide, and improve academic performance in today's classrooms.

Appendix A: PLATO Learning Item Writing Guidelines²

Goal in item writing

The most important goal in test item writing is to create a series of items that are the most direct measure of the intended learning outcome. There are barriers to this goal, including ambiguous statements, excessive words, difficult vocabulary, unclear instructions, and bias. The first step in getting around barriers is to select the appropriate item type for the objective.

Selecting the appropriate item type

PLATO Learning provides a reference table to help item writers determine the best item type.

Item Type Table

Multiple Choice	True/False	Matching	Grid Response
<ul style="list-style-type: none"> ■ Useful when other item types can't be used ■ Best answer type is most difficult for students ■ Material need not be homogeneous ■ More choices, more reliability ■ Limited to learning outcomes that are verbal (as opposed to natural situations such as performance assessments) ■ Creating good distractors is difficult 	<ul style="list-style-type: none"> ■ Most useful when there are only two possible alternatives, and when distinguishing fact from opinion, cause from effect, etc. ■ Limited reliability due to guessing ■ Useful for knowledge only (not higher-level thinking) 	<ul style="list-style-type: none"> ■ When learning outcome emphasizes relationship between two things ■ Limited to factual information ■ Material must be homogeneous 	<ul style="list-style-type: none"> ■ Most useful for factual information ■ Used when answers are numerical ■ Limited due to structure

Principles and Guidelines

For each item type, there are principles to follow to help reduce ambiguity and ensure that the item matches the intended learning outcome. PLATO Learning uses these principles as guidelines for writing items.

Guidelines for All Item Types

- Use novel material to test higher level learning.
- Avoid trivial content. Base each item on an important outcome for the student to learn. One way to differentiate trivial information from information that is not trivial is to answer this question: Is this a fact that is virtually meaningless to most students? This method won't work in every case, of course, as there are cases to be made for memorizing terms, etc. But it is a good rule of thumb.

² Haladyna, Thomas M. *Developing and Validating Multiple-Choice Test Items*. Linn, Robert L. and M. David Miller. *Measurement and Assessment in Teaching*.

- Avoid irrelevant information (window dressing). An item that contains irrelevant information uses words, phrases, or sentences that have nothing to do with the problem stated in the stem. The intention is usually to add substance to the item, but generally window dressing is not needed.
- Avoid trick items. Trick items are intended to deceive the test taker into choosing a distracter instead of the correct answer. A negative aspect of trick items is that if they are frequent enough, they build an attitude by the test taker of distrust and potential lack of respect for the testing process. There are two types of trick items: items deliberately intended by the writer and items that accidentally trick test takers.

Here are seven types of items that students perceive as tricky:

1. The item writer's intention appears to be to deceive, confuse, or mislead test takers.
2. The item includes trivial content.
3. The discrimination among items is too fine.
4. The item has window dressing that is irrelevant to the problem.
5. More than one correct answer is possible.
6. Principles are presented in ways that the student has not learned.
7. Items are so highly ambiguous that even the best students have no idea what the correct answer is.

- Readability. Ensure that vocabulary and sentence structures are appropriate for the intended grade level. In particular, avoid long, complex sentence structures in stems that can hinder comprehension (see the example below).
- Vary the location of the correct answer. The correct answer should appear an equal number of times in each answer position.

Multiple-Choice Items Guidelines³

- Include the central idea in the stem, not in the answer options.
- Avoid negatively stated stems.
- Items should include a minimum of two and a maximum of six answer options, but aim for three to four answer options whenever possible.
- All answer options must be plausible and realistic. Ways to make answer options plausible include the following:
 - a) Using students' most common errors
 - b) Using important sounding words (e.g., significant, accurate)
 - c) Using incorrect answers that are likely to result from student misunderstanding or carelessness (e.g., student forgets to convert from feet to yards)

³Taken from the PLATO Learning Test Packs Item Writing Guide, originated from Linn, Robert and M. David Miller, Measurement and Assessment in Teaching.

-
- All answer options must be similar in length and grammatical structure. In particular, do not make the correct response different grammatically, in length, etc.
 - All answer options must be in logical order (either alphabetical or numerical).
 - Avoid using *none of the above* or *all of the above* as an answer option. When students give these responses, it's not clear whether they know the correct answer.
 - Avoid repeating words from the question stem in the answer options. For example, if you use the word *experiment* in the stem, using that same word in an answer option may lead a student to select that response.

True/False Guidelines

- Avoid negatively stated stems.
- Do not use double negatives in stems.
- Do not use definitive words such as *all*, *only*, *none*, *always*, and *never* that could lead students to answer *false*.
- Do not use uncertain words such as *usually*, *might*, *can*, and *may* that could lead students to answer *true*.
- Use precise wording (*12 years old*, *more than 6 feet tall*, and *60,000 people*) rather than vague or qualitative wording (*young*, *tall*, *many*).
- Statements must be completely true or completely false. However, in an attempt to make sure statements are perfectly true or false, items are sometimes created that have little significance from a learning standpoint. Make sure stems are meaningful.
- Use simple, easy-to-follow statements that are free of trivial information.
- Unless measuring cause-effect relationships, avoid including two ideas in one statement.
- Avoid opinions unless they are attributed to a source.
- True statements and false statements should be approximately equal in length.
- The number of true statements and false statements should be approximately equal. The emphasis here is “approximately equal”. One guideline suggests about 40% true statements and about 60% false.

Matching Guidelines

- Matching questions are most appropriate for testing the student’s ability to recognize the following:

People and their achievements	Plants/animals and their classification
Dates and events	Principles and illustrations
Terms and definitions	Objects and names of objects
Rules and examples	Causes and their effects
Symbols and concepts	Problems and solutions
Authors and book titles	Parts and their names
Machines and uses	Parts and their functions
- Use only homogeneous material in a single matching exercise. This is the most important rule of matching construction, but also the most violated.
- Include an unequal number of responses and premises, and instruct the student that responses may be used once, more than once, or not at all. This approach decreases the likelihood of guessing and makes all the responses eligible for each premise.
- Keep the list of items to be matched brief and place the shorter items on the right.
- Arrange the list of responses in logical order; place words in alphabetical order and numbers in sequence.
- Give directions. Although matching is rather obvious, it’s important to clearly state the directions (what is being matched to what) to avoid ambiguity and confusion. (Example: Match the correct date to each event.)

Grid Response Guidelines

Grid response (sometimes called “grid in”) is a short-answer item type. Short-answer item types can be answered by a word, phrase, number, or symbol. Grid response items must be answered using numbers or symbols.

Grid response questions should be used only to:

- measure understanding of specific facts;
- measure interpretations of data;
- measure performance of computational procedures.

Grid response questions are especially useful for measuring recall of information.

They can be used:

- for any question with a response that can be expressed using the slash symbol, the decimal/period symbol, and/or the numbers 0-9;
- when allowing students to “test” all the multiple choice answers to see which one works isn’t desirable;
- when it’s hard to have more than one plausible distractor;

- when it's not ideal for a learner to choose the best answer from multiple choices;
- with tests for grade 4 and above;
- on 15 percent or fewer items on a test version, as long as the test has at least four grid response items per version.

Grid response questions should be:

- grouped together within the structure of a test;
- used when there is only one way to answer the question;
- used when the answer isn't trivial (e.g., use for important dates such as the date the Declaration of Independence was signed or the date Japan bombed Pearl Harbor).
- used when the item type won't get in the way of measuring the objective.

Other guidelines:

- Word the item so the required answer is brief and specific.
Avoid: When was George Washington born?
Better: What year was George Washington born?
- Do not take statements directly from textbooks. Textbook statements are often too general and ambiguous to be good short answer items.
- Use a direct question when possible as opposed to an incomplete statement.
Poor: John Glenn made his first orbital flight around Earth in _____.
Better: When did John Glenn make his first orbital flight around Earth?
Best: In what year did John Glenn make his first orbital flight around Earth?
- If the answer is to be expressed in numerical units, indicate the type wanted.
Poor: If oranges weigh $5 \frac{2}{3}$ ounces each, how much will a dozen oranges weigh? (4 pounds, 4 ounces)
Better: If oranges weigh $5 \frac{2}{3}$ ounces each, how much will a dozen oranges weigh? ____pounds, ____ ounces. (4 pounds, 4 ounces)
- Don't use extraneous columns with lower grades. Although ten columns are available, use only four columns for the numbers in a year (e.g., 1776). Also, provide a dollar or cents sign in the appropriate column.
- The item should have enough columns to accommodate the range of plausible answers. In other words, for computations, don't give away the number of digits of the response by having the exact number of columns needed for the right answer. If a reasonable wrong answer might have more columns than the correct answer, make sure there are enough columns for that plausible wrong answer to be entered.

PLATO Learning Style Guides

PLATO Learning provides item writers with several robust style guides, including a writing style guide for each of the content areas: math, English/language arts, science, and social studies. In addition, technical and graphic style guides were developed for use with the tool used to create online test items.

Targeted Trainings

Targeted trainings are developed around each subject area (math, English/language arts, science, and social studies) to guide writers in creating new items. First, a set of sample items is written and peer reviewed by PLATO Learning subject matter experts and instructional designers. Using the quality results of the review as a needs assessment, subject area leads then hold training meetings to identify the training sessions to develop. These targeted trainings include topics such as writing grade-level appropriate items, best practices in assessment writing, PLATO Learning's acceptance criteria for quality assurance, and avoiding bias in item writing.

Item Evaluation: Acceptance Criteria

In addition, PLATO Learning uses the item writing guidelines to create acceptance criteria. The acceptance criteria is used to evaluate every item written to determine if it meets PLATO Learning's expected quality level. Subject area teams work together to peer review items, using the acceptance criteria and item writing guidelines to evaluate the quality of the item. If items do not meet the acceptance criteria, the item is not used on a PLATO Learning test. In addition, the acceptance criteria forms are tracked using metrics which help PLATO Learning developers make decisions about needed trainings.

PLATO Learning Acceptance Criteria

	Showstoppers	Any “No” results in a rejection	Yes	No	Reject
Correct answer	1.1	The right answer is accurate & correct for each item.			
	1.2	There is only 1 correct answer for each item.			
Grade level	1.3	The stems are grade level appropriate.			
	1.4	The answer options are grade level appropriate. (difficulty)			
	1.5	The answer options are grade level appropriate. (readability) [Item should not be above grade level, 1 grade level below is acceptable]			
	1.6	The answer options are homogeneous enough that it doesn't give away correct answer.			
Objective	1.7	The item meets the objective.			
Content	1.8	The content is correct in the stem and answer options.			
Bias	1.9	The item is free of bias or subjectiveness.			
	General	Internally tracked, does not result in rejection	Yes	No	Reject
Assessment best practice	2.1	The stem is a complete sentence and is either a question or an imperative statement.			
	2.2	The answer options are free of unnecessary words.			
	2.3	The answer options are homogeneous in nature (in length, verbs, and wording).			
	2.4	All answer items are plausible (likely).			
	2.5	<i>None of the above</i> and <i>All of the above</i> have been avoided.			
	2.6	Instructions are clear.			
	2.7	The stem is meaningful by itself.			
	2.8	The items are measuring non-trivial content.			
	2.9	Negatively stated items have been avoided.			
Editorial	2.10	There are no spelling or grammar errors.			
	2.11	Answer options are in alphabetical order.			
	2.12	Answer options are in numerical order.			

Appendix B: Readability Measures

Determining readability is not a precise metric or science. Reading experts continue to debate and evolve the formulas of various readability scales. The following are some common criteria used to level text. Instructions and other text in math, social studies, and science will not be given a readability level.

The following are some computerized materials used to level text. While every scale listed below is used for every passage, PLATO Learning uses a set of criteria in order to determine appropriate readability.

Computerized leveling *(for materials from elementary through adult)*

- **Word frequency:** text is compared to a corpus, a list of millions of words usually drawn from educational materials; software measures how often each word being compared appears in the corpus.
- **Sentence length:** the average number of words per sentence
- Number of syllables per word
- Length of overall text
- Percent of familiar words versus content words

1. Readability Plus — PLATO Learning’s content development and editorial departments primarily use Micro Power and Light’s Readability Plus (Formula Set 2) software to determine readability scales.

The Readability Plus scales that PLATO Learning uses are the Fry Graph as well as the Spache, Powers-Sumner-Kearl, Dale-Chall, and Flesch scales. These readability formulas are listed below with a brief description.

Lower Elementary (Grades 2–4)

- **Fry Graph**—This readability formula is designed for all levels of reading matter, from elementary through college. It is intended for instructional materials, and it assumes that the teacher will explain some of the material. Due to this assumption, the Fry Graph accepts 65–75 percent comprehension as adequate.
- **Spache**—This readability formula compares the selected text to a list of standard vocabulary words. It applies to materials intended for primary through fourth grade.
- **Powers-Sumner-Kearl**—This formula-based readability scale is designed for primary materials with a ceiling of seventh grade.

Upper Elementary and Middle School (Grades 5–8)

- **Fry Graph**—See previous explanation.
- **Dale-Chal**—This readability formula is highly regarded by experts in the field of education and the subfield of Reading. The Dale-Chall formula has credibility because it does not measure word difficulty solely by length, as most others do. Instead, it checks the text against its list of “familiar” words

that the students should know; words that are deemed difficult are not included in the Dale-Chall list. This readability formula also factors in sentence length. It is designed for upper elementary and secondary materials only.

- **Flesch**—This grade-level formula relies on numbers of syllables, words, and sentences. It is intended for secondary materials. (Please note that PLATO uses the Flesch scale from the Micro Power and Light software to provide readability statistics for passages in the PLATO database. PLATO Learning does not use the Microsoft Word Flesch-Kincaid readability tool. For this reason, variations in Flesch readability results may appear when passages are cut and pasted from the PLATO Learning passage database into a Microsoft Word document.)

High School (Grades 9–11)

- **Flesch**—This grade-level formula relies on numbers of syllables, words, and sentences. It is intended for secondary materials. (Please note that PLATO Learning uses the Flesch scale from the Micro Power and Light software to provide readability statistics for passages in the PLATO database. PLATO Learning does not use the Microsoft Word Flesch-Kincaid readability tool. For this reason, variations in Flesch readability results may appear when passages are cut and pasted from the PLATO Learning passage database into a Microsoft Word document.)

While the Micro Power and Light software is effective in providing readability scales for lengthier passages, the software is not intended for use on individual items. These quantitative methods also do not take into account the interest or maturity level of the passage, and these considerations affect a student’s ability to comprehend a passage. Additionally, these scores are estimates, not absolutes. Formulas cannot detect and measure abstract ideas or difficult concepts. They also cannot measure “implications” such as cause-and-effect relationships, quality of writing, clarity of style, or textual organization. Finally, these programs do not exclude proper names or other nouns that, while an issue for readability, might be overlooked by the skilled reader. For these reasons, PLATO Learning also relies on the professional judgment of its content specialists when assessing the readability of passages and items.

2. EDL Core Vocabularies

3. Expert Judgment—PLATO Learning relies not only on established readability formulas and core word lists, but also on the judgment of its content specialists for determining appropriate readability of passages and items. PLATO Learning utilizes the experience of its educational specialists who are former teachers and who have an understanding of students’ critical and creative thinking skills at specific grade levels. PLATO Learning’s educational specialists also consider a passage’s unity, structure, degree of interest and appeal, and sophistication of concepts and themes to determine its appropriateness.

4. Lexile Levels—PLATO Learning uses Lexile, a key readability tool that leading assessment developers of educational content consider appropriate for elementary and middle school levels. Lexile levels are provided through MetaMetrics. MetaMetrics develops scientifically based measures of student achievement that link assessment with instruction, foster better educational practices, and improve learning by matching students to materials that meet and challenge their abilities. Lexile levels were developed by researchers at the National Institute of Child Health and Human Development. The text is analyzed by Lexile software and given a measure from below zero to 2000 based on the number of words per sentence and word frequency. Students are given a measure on the same scale based on standardized tests. Students are then matched with materials that have this same measure. Lexile levels are used for elementary through high school grades and are used by many textbook, trade book, and school library publishers.

Hand Leveling by Educators

- Used mainly for materials in elementary guided reading and early intervention programs.
- **Text length:** number of pages, number of words, and number of lines on a page
- **Layout:** font size, space between words and lines, positioning of text on a page
- **Structure and Organization:** difficulty of plot, amount of repetition
- **Illustrations:** how much they support the text
- **Words:** frequency, number of syllables, number of content words
- **Phrases and Sentences:** how complex they are
- **Literary Features:** complexity of ideas, use of flashbacks or metaphors
- **Content and Theme:** how familiar the content is; how directly it relates to the reader

PLATO Learning uses Lexile as a key readability tool that leading assessment developers of educational content consider appropriate for elementary and middle school levels. Where Lexile levels are listed, they have been provided by MetaMetrics. MetaMetrics develops scientifically based measures of student achievement that link assessment with instruction, foster better educational practices, and improve learning by matching students to materials that meet and challenge their abilities. Lexile levels were developed by researchers at the National Institute of Child Health and Human Development. The text is analyzed by Lexile software and given a measure from below zero to 2000 based on the number of words per sentence and word frequency. Students are given a measure on the same scale based on standardized tests. Students are then matched with materials that have this same measure. Many textbook, trade book and school library publishers list Lexile levels in their products.

Appendix C: Comparability of Test Versions

These comparability guidelines were created to accomplish the following goals:

- Prevent information gaps about the issue of comparability
- Document rules for comparable test and item creation for PLATO Learning's assessment product(s)

Definitions

Comparable items are written to reflect assessment benchmarks and reporting categories at the objective level. These items consistently represent the same features, parameters, and attributes. Comparable items measure equivalent knowledge and skills at the same cognitive level and with the same degree of difficulty.

Comparable passages are written according to the same readability and reading ease scales. Comparable passages represent identical genres, length, and textual attributes. Primary genres are expository, narrative, and poetic texts. Comparability between passages includes appropriate considerations of subgenre, topic, stylistics, theme, characterization, and setting.

Comparable tests are written as alternate and comparable forms of assessments that align with blueprint specifications. Alternate forms incorporate comparable items with equivalent format, item count, and distribution according to designated objectives and reporting categories.

Comparability of Test Items

When judging test items to make sure the test versions are comparable, there are four major areas the Test Packs development team addresses:

- Cognitive level
- Readability
- Plausibility of answer options
- Visuals

If items are comparable in these four areas within a reasonable margin, then the items have been considered comparable for the purposes of generating comparable test versions. To make these judgments, developers use a series of guidelines for each area. The guidelines are not enforced word-for-word in all cases. Instead, they describe an ideal level of comparability at a micro-level. The characteristics of an item as it relates to each of these guidelines is then factored into higher level judgments about the four major areas of comparability. In addition, the four areas have relative levels of importance. For example, more importance is placed on cognitive level and readability than on visuals when making a comparability judgment about an item.

The following characteristics apply to a test as a whole, rather than to individual test items.

Test Characteristic	PLATO Learning Rule or Procedure
Number of Items	Each test should have same number of items on all versions.
Reporting Categories	Each test should have same reporting categories and same number of items within each category as previous versions.
Length	Length of test should be comparable to previous versions.

The table that follows describes several characteristics of test items and PLATO Learning's procedure for evaluating each characteristic in order to ensure that each new item is comparable to peer items on a test.

Guidelines for Test Item Comparability

Cognitive Level	
Difficulty Level	New item should have same difficulty as original – at grade level consistent with item. (Note: Grade-level difficulty is a different issue from individual item's level of difficulty. Item "A" and item "B" may both be acceptable in terms of difficulty for the grade, but item "A" might be a more difficult item than item "B.")
Reasoning Process	New item should have same number and type of cognitive steps required as original.
Number of Steps	New item should have same number of steps.
Cognitive Task	New item has to match other item.
HOTS Level	New item must match original item.
Readability	
Language, Vocabulary	New item should have same difficulty of language arts/reading level, consistent with other items. PLATO Learning uses the Children's Writers Word Book and the EDL Core Vocabulary list to help determine the difficulty of vocabulary. Item statement must not require a higher or lower level of language arts/reading comprehension.
Length	New item should be approximately the same number of words within + or - 25% for the stem and the answer choices.
Context/Context Boundaries	Context of item is mostly based on information within the passage. This is where the issue of passage comparability comes into play. For example, if item from original passage addresses internal conflict and second passage has external conflict, context of the items associated with second passage will be slightly different. Passage attributes must be followed to produce comparable passages, ensuring that context issues with items are not a problem.
Plausibility of Answer Options	
Answer Option Rationale	Correct answer and two or more incorrect options must have same rationale as original item (i.e., is one answer abstract, one specific, one general?).
Answer Option Positioning	Position of correct answer (either A, B, C, or D) can vary between original and new item.
Visuals	
Graphics (Inclusion of)	If original item has graphic, then comparable item must have graphic. Purpose of graphic must be the same.
Graphics (Originality of)	New item must have graphic different from original graphic.
Visual Impact/Layout	Orientation should not impact level of difficulty. Includes proximity of clue to the question. Note: Using an excerpt from the text may actually make the item a few "steps" easier.

References

- Bernhardt, Victoria L. 2003. *Using data to improve student learning in elementary schools*. Larchmont, N.Y.: Eye on Education.
- Doran, Harold C. 2003. Adding value to accountability. *Educational Leadership* 61 (3): 55–59.
- Finn, Chester E., Jr., Liam Julian, and Michael J. Petrilli. 2006. *The state of state standards, 2006*. Washington, D.C.: Thomas B. Fordham Foundation.
- Gandal, Matthew, and Laura McGiffert. 2003. The power of testing. *Educational Leadership* 60 (5):39–42.
- Gronlund, N.E. 1993. *How to make achievement tests and assessments*. Boston: Allyn and Bacon.
- Gross, Paul, Ursual Goodenough, Lawrence Lerner, Susan Haack, Martha Schwartz, Richard Schwartz, and Chester Finn Jr. 2005. *The state of state science standards, 2005*. Washington, D.C.: Thomas B. Fordham Foundation.
- Guskey, Thomas R. 2003. How classroom assessments improve learning. *Educational Leadership* 60 (5):7–11.
- Haladyna, Thomas M. 2004. *Developing and validating multiple-choice test items*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Halyadyna, Thomas M. 1997. *Writing test items to evaluate higher order thinking*. Boston: Allyn, and Bacon.
- Jerald, Craig. 2003. Beyond the rock and the hard place. *Educational Leadership* 61 (3):12–16.
- Klein, David, Bastiaan Braams, Thomas Parker, William Quirk, Wilfried Schmid, W. Stephen Wilson, Chester E. Finn Jr., Justin Torres, Lawrence Braden, and Ralph Raimi. 2005. *The state of state math standards, 2005*. Washington, D.C.: Thomas B. Fordham Foundation.
- Linn, Robert L., and David M. Miller. 2005. *Measurement and assessment in teaching*. 3rd ed. Upper Saddle River, N.J.: Pearson.
- Mead, Walter, Chester Finn Jr., and Martin Davis Jr. 2006. *The state of world history standards, 2006*. Washington, D.C.: Thomas B. Fordham Foundation.
- National Research Council. 2000. How people learn: *Brain, mind, experience, and school*. Washington, D.C.: National Academy Press.
- National Research Council. 2001. *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.
- Osterlind, S.J. 1997. *Constructing test items: Multiple-choice, constructed response, performance, and other formats*. Boston: Kluwer Academic Publishers.
- Rettig, Michael D., Laurie McCullough, Karen Santos, and Chuck Watson. 2003. A blueprint for increasing student achievement. *Educational Leadership* 61 (3):71–76.
- Sharkey, Nancy S., and Richard J. Murnane. 2003. Learning from Student Assessment Results. *Educational Leadership* 61 (3): 71–81.
- Stotsky, Sandar, and Chester E. Finn Jr. *The state of state English standards, 2005*. Washington, D.C.: Thomas B. Fordham Foundation.