ED 469 376 TM 034 526

AUTHOR	Foshay, Rob
TITLE	Choosing the Right Testing Option in PLATO Courseware. PLATO Technical Paper.
INSTITUTION	PLATO Learning, Inc., Bloomington, MN.
REPORT NO	PLATO-TP-13
PUB DATE	2002-08-00
NOTE	60p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	Achievement Tests; Computer Software; Curriculum; Mathematics; Needs Assessment; *Reading; *Selection; Student Placement; Test Coaching; *Test Use; *Writing (Composition)

ABSTRACT

There are a number of issues to consider in choosing tests, including alignment of tests and standards, the integration of tests with curriculum and instruction, the quality of the tests, and a clear definition of the purpose of the test. To address the various needs reflected by these issues, PLATO Learning, Inc., offers two curriculum-wide testing systems: NetSchools "Orion" GATE and PLATO (registered) LINK. For practice in preparation for high-stakes tests, PLATO Learning offers the Simulated Tests in mathematics, reading, and writing. To support needs for placement, progress control, and cumulative testing when using PLATO (registered) courseware, PLATO Learning offers the FASTRACK and Skills Inventory systems, module mastery tests, and course-level assessments. Each of these systems has different characteristics and is designed to serve different needs. Choosing among them involves answering 12 key questions about testing needs, which are provided in this document. (Author/SLD)

PLATO®

Technical Paper #13

August, 2002

Choosing the Right Testing Option in PLATO Courseware

Rob Foshay, Ph.D.
Vice President

Instructional Design and Cognitive Learning

PLATO Learning, Inc.
10801 Nesbitt Avenue South
Bloomington, MN 55437
(800) 869-2000
http://www.plato.com
author's e-mail:
rfoshay@plato.com

ABSTRACT

Any effective educational enterprise must include measurement of its learning outcomes, before, during and after instruction. Standards and accountability requirements place even more emphasis on assessment. But there are a number of issues to consider when choosing tests. Main ones include:

- Alignment of tests and standards
- Integration of tests with curriculum and instruction
- Quality of tests, including validity issues such as detailed alignment with standards, as well as reliability
- Clear definition of the purpose of the test, including high vs. low stakes, two types of placement testing needs, progress control, and two types of cumulative post-tests.

To address the various needs emerging from these issues, PLATO Learning, Inc. offers two curriculum-wide testing systems, NetSchools *Orion* GATE, and PLATO© LINK. For practice in preparation for high-stakes tests, PLATO Learning offers the Simulated Tests in math, reading and writing. To support needs for placement, progress control and cumulative testing when using PLATO© courseware, PLATO Learning offers the FASTRACK and Skills Inventory systems, module mastery tests, and course-level assessments. Each of these systems has different characteristics and is designed to serve different needs. Choosing among them involves answering 12 key questions about your testing needs.

Table of Contents

Part

1

Introduction

Any effective educational enterprise must include measurement of its learning outcomes, before, during and after instruction. Standards and accountability requirements place even more emphasis on assessment. One result is that an increasing number of tests have been layered onto the curriculum, often without enough attention to how well they reflect standards or how well they measure. Furthermore, data obtained from the tests are often so general and so slow in coming that they are of little use to teachers and administrators who need to make decisions about individual learners, classes and schools. The net result can be highly misleading information about program effectiveness: *the wrong test, reported at the wrong time, gives* ***disinformation*** *to educators, policy makers and the community.*

In response to this problem, PLATO Learning, Inc. has recently expanded the capabilities of the PLATO© family of technologies for testing. Our goal is to provide educators at all levels with high-quality, standards-aligned, competency-based, online testing systems, with immediate online reporting, for the full range of low-stakes testing needs. The testing systems support both the PLATO© instructional systems and the full core curriculum. Many of the testing systems are customizable, and some may be used independently of the PLATO courseware if desired.

This technical paper will first review key issues in testing and discuss their relationship to effective implementation of standards under *No Child Left Behind.* Then it will provide an overview of each testing capability provided by the PLATO© technologies. Finally, a guide to choosing among the testing options will help you choose among the options PLATO Learning provides.

Part

2

Using Tests to Support Effective Standards Implementation

In this part, we'll first discuss five common issues in standards implementation which surround use of tests. Then we'll discuss the types of tests that are needed to implement standards. We will then briefly summarize the testing requirements of *No Child Left Behind.* Finally, we'll discuss issues of reliability and validity as they affect the types of tests needed for standards implementation, and provide an overview of validity and reliability procedures used for PLATO© tests.

5 Common Testing Issues in Standards Reform

Tests are by far the most common means of assessment in education, but their very familiarity often leads educators to overlook issues in test design that take on particular significance when implementing a standards-based system for accountability. These issues concern:

- Confusion over test types
- Disconnected curricula and tests
- Technically poor tests
- "The tail wags the dog" syndrome
- Testing overload

We'll discuss each of these in turn.

Confusion over test types

Educators are most familiar with *norm-referenced* tests (often called "standardized tests"), yet adequate measurement of standards requires *criterion-referenced* tests. The distinction between the two is not widely understood.

A *norm-referenced* test is designed to compare achievement of each learner to a reference group, such as a national sample of students at the same grade level. Questions on a norm-referenced test are chosen because they are of moderate difficulty for students at that grade level: if a question is too hard or too easy, it is eliminated because it doesn't do a good job of classifying students. The content of a norm-referenced test is determined by a *domain specification* which carefully defines the boundaries of the content area to be measured. For example, grade levels in reading are norm-referenced. "Grading on the curve" is a norm-referenced practice. Tests such as the *Iowa Tests of Basic Skills* and the *Stanford Achievement Test* are norm-referenced.

A *criterion-referenced* test is designed to map how well learners can perform (or understand) a particular benchmark for a standard. The difficulty of the questions included is determined by the benchmarks to which they correspond: each question will be only as easy or hard as is needed to properly measure the benchmark. Students who have fully mastered the standards should find the test easy. In the criterion-referenced world of standards, if the system is working, everyone should get an "A." Even better than a letter grade, however, is a checklist showing which benchmarks they have (and have not) attained – in effect, a separate "pass/fail" decision on each benchmark.

The content of a criterion-referenced test is usually *competency-based*, meaning that it takes as its content map the standards and benchmarks to be measured. Rather than mapping the "boundary" of the content domain as a norm-referenced test does, a criterion-referenced test maps the whole of the content area. State standards tests are usually competency-based and criterion referenced, and Federal policy under *No Child Left Behind* encourages states now using norm-referenced tests to transition to competency-based ones.(Marzano and Mid-Continent Regional Educational Lab. Aurora CO. 1998)

You can't simply look at a test and tell whether it is norm-referenced or competency-based. The question types on the two tests usually are the same, and the items are scored in the same way. There might be differences in details of what is tested or in item difficulty which wouldn't be apparent on a casual inspection. The basic distinction really is with the interpretation of the score: a norm-referenced test classifies learners relative to other learners; a competency-based test makes "yes or no" decisions on whether the learner has mastered each competency tested. (Linn, National Council on Measurement in Education. et al. 1993)

Disconnected curricula and tests

Truly implementing state curriculum standards in the daily practice of the classroom is a tremendous challenge. Most educators confront a *disconnected* curriculum structure: National standards (such NCTM) and tests (such as the National Assessment of Educational Progress, or the Scholastic Aptitude Test) show major discrepancies between each other and with the various state standards. Often, the state standards and tests do not align well with each other(Marzano and Kendall 1996). Furthermore, the standards and benchmarks themselves are often of poor technical quality (Kendall 2001), and educators find them difficult to interpret at the level of detail needed for daily lesson planning and testing. The sheer volume of standards is itself an issue. National and state content standards would require students to master one benchmark per day in every subject – an unrealistic goal (Marzano and Kendall 1996), so schools must choose what standards and benchmarks to implement.

In this climate, it is scarcely surprising that simply figuring out what to teach and test is extremely difficult. Most teachers, confronted with a thick binder of standards which may suffer from these weaknesses, find it impossible to actively use them as a guide to daily teaching and testing. The result is that what happens in the classroom often has only an indirect relationship to standards.

Generations of educational practice often lead well-meaning professionals to focus on the wrong things. The essence of the standards movement is to define success in terms of learning outcomes (what the students can do) rather than delivery of instruction (what the teachers do). Yet, many standards documents focus on *content* rather than *performance*. Furthermore, teachers often develop a repertoire of activities which seem to work well with their learners, and they may be reluctant to change, even if the activities have little relationship to curriculum standards. When teachers develop their tests based on these activities, the influence of standards is lost.

Administrative practices also reinforce the disconnections. Schools usually standardize instruction, rather than learning: everyone gets the same number of "contact hours" in each subject (leading to the familiar "bell-shaped curve" of achievement), rather than creating a system which does whatever it takes so everyone reaches the same learning outcomes (leading to a "bell shaped curve" of instructional time). Even the familiar Carnegie Unit is defined in this way. This virtually insures that mastery of standards by all learners will not occur. Grading systems are based on the norm-referenced idea of "grading on the curve" rather than the competency-based framework of standards. Even when tests are standards-referenced, there are often delays of months in reporting results, and the reports often lack the detail needed to plan interventions with particular students.

The result is a system which in which the teaching practices, the content, the class time, the teaching, the administrative practices and the tests are all disconnected from each other and from curriculum standards.

Technically Poor Tests

Researchers have shown that many state tests are of poor technical quality (Marzano and Kendall 1996). They often have only a modest correspondence to the standards they are intended to test, and they may not have gone through a sufficiently rigorous item development process to justify the tests' use in high-stakes situations such as determining eligibility to graduate. Some states have adopted norm-referenced standardized tests, rather than developing their own. However, a norm-referenced test is not an adequate measure in the competency-based world of standards, and Federal policy under *No Child Left Behind* is to move toward competency-based tests.

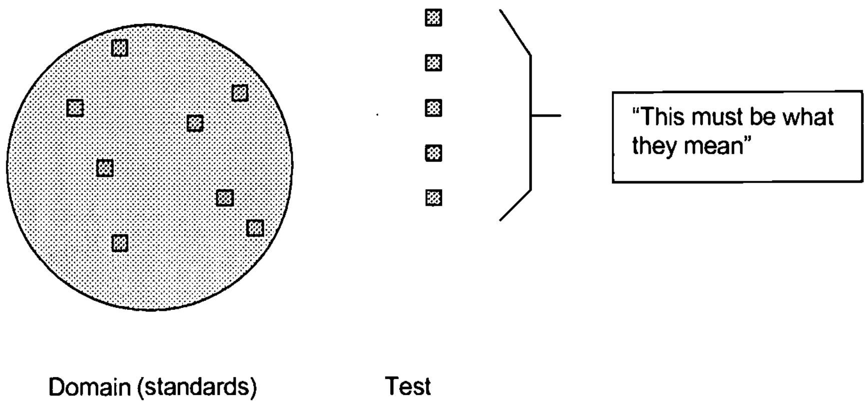
District and classroom tests typically are for low-stakes purposes, and the development cost and effort can be lower than for the high-stakes state tests. However, teacher-made tests are of notoriously uneven quality, often with overemphasis on questions which are low on Bloom's *Taxonomy of the Cognitive Domain*, and which exhibit many stylistic flaws which depress test quality. Since, as argued above, these tests are often disconnected from standards as well, the result is that the "first line" of measurement of learning in the classroom is often highly misleading. It's little wonder that administrators are vulnerable to "nasty surprises" when students who have been testing well all year on the local measures suddenly have difficulty on the high-stakes state tests.

It is also clear that not all benchmarks can be adequately measured with multiple-choice questions (Marzano and Kendall 1996). Educators must be clear about what can – and cannot – be measured by multiple-choice tests, and a complete measurement of standards must include the more labor-intensive performance-based measures in a portfolio evaluation approach, even though reliability of these measures is often lower than a well-designed multiple-choice test.

"The Tail Wags the Dog" Syndrome

The problems with disconnected curricula and poor tests lead to a further difficulty: "teaching to the test" – the *wrong* way. The problem starts with the principle that since it's impractical to test for *all* the learning outcomes referenced in the standards, the state tests must use a *domain sampling* procedure. In other words, each test is a random sample of just a few of the many questions which correspond to the standards. In Fig. 1 (next page), the domain (described by the standards) is represented by the circle on the left, and the random sample of items are the squares within. These are then assembled into a test, as shown by the column of squares on the right.

The problem comes when teachers, seeking guidance on what to teach, look at the test (the domain sample) rather than the curriculum standards (the domain). They conclude that "this must be what they mean" by the standards, and follow the test. However, since the test is only a domain sample, the disconnect with the standards is still a problem. Furthermore, when a new form of the test is used (perhaps in the next semester), it has a different domain sample, and the teacher will complain that "they changed the test," thus betraying both teacher and students.



• Fig. 1: Teaching to the test

"Teaching to the test" would be acceptable – if the test were a perfect and complete measure of the standards (the domain). But most tests are domain samples, and no test is a perfect measure of its standards. Therefore, the goal should be to teach to the standards, not to the test.

Testing Overload

Standards and accountability reforms have often been treated as additive, rather than displacing old policies. One result of this trend has been that a typical curriculum has many "layers" of testing, often including national tests such as the NAEP, SAT, ACT, state standards tests, norm-referenced tests used by the district as pre- and post-tests, district-mandated final exams, various diagnostic and placement tests, departmental unit tests, and each teacher's own unit tests and quizzes. It's little wonder that teachers complain that so much time is occupied by testing that there's scant room left for teaching. The problem is exacerbated by the disconnections between the tests, as discussed above. It's nearly impossible for a teacher to figure out which tests are providing data which actually relates to standards.

Fixing a "no-win" situation

The issues of test quality, disconnects with the curriculum, the "tail wagging the dog," and test overload often combine to produce a no-win situation for teachers. Faced with uninterpretable standards, misaligned tests of variable quality, and the imperative to improve test scores, teachers often become frustrated and even cynical about testing in general: they feel they are in a no-win situation.

Fixing the problem is at the core of successful standards implementation:

- Schools should insist that their state standards be technically sound, with clear definitions of domains, benchmarks and performance levels—in numbers realistic to teach in the time available. Where this is not the case, schools will need to develop their own "interpretation" of their state's standards.

- Schools and state policy makers should make sure all assessments (high and low-stakes) are aligned to their standards and benchmarks – and eliminate those tests which are not.

- Schools should define what to teach in terms of competency-based benchmarks and objectives, not in terms of tests.

- Leaders should focus attention on student attainment of the standards, not simply on quality of teaching.

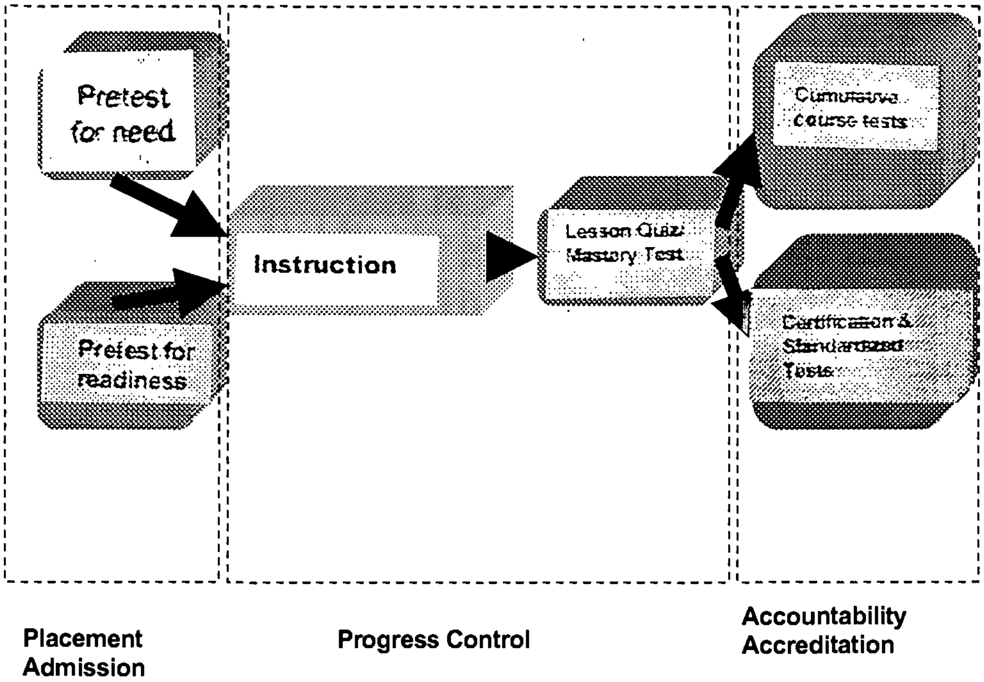
- Leaders should make sure high- and low-stakes assessments are of the appropriate types, that they are competency-based and criterion-referenced.

- Within the practical limits of assessment length and cost, leaders should make sure *all* tests are of appropriate quality for their purpose.

- Schools should arrange for timely and detailed (disaggregated) reports from all tests, so they can be used for data-based decision-making.

Types of Tests

Five types of tests are commonly used in schools. They differ by purpose, content, and quality/cost/length requirements. The test types are shown in Fig. 2. We'll discuss each type separately.



• Fig. 2: Types of Tests

Lesson Quiz/Mastery Tests

The most familiar kind of test, these tests are usually given in close proximity to instruction – often at the end of the lesson (as is the case with PLATO courseware). The purpose of the test is to provide the learners with quick feedback on how well they understand what was just studied, and to provide the instructor with rapid information which can be used to decide what to do next: reteach, or go on to the next lesson. In mastery learning, these tests are used to regulate progress through the learning sequence.

Content of the test should reflect the lesson's terminal objective(s). This close linking of instruction and test is especially important here, because the tests should be used for data-based decision-making to guide each individual learner.

In general, in individualized instruction, it is better to give many short "testlets," tightly linked to instruction, rather than a few longer tests. The "testlets" can be short quizzes, which only focus on a single terminal objective, and which are very short (research on test length suggests that 8 questions per objective provides the best reliability). This allows for immediate identification of learning problems and immediate intervention.

It is important for these tests to be a highly valid measure of what is taught in the instruction. If the instruction is in turn defined by standards, then the lesson quiz will be a valid measure of the corresponding benchmark or objective.

This is a low-stakes test, since the only consequence of measurement error would be to waste a small amount of the learner's time (by reteaching or by going on inappropriately). Therefore, reliability can be moderate, and the test cost and test length can be modest.

While these tests are commonly graded by percentage and used as a basis for letter grades, in the competency-based, criterion-referenced world of standards, it makes more sense for these tests to be graded as "pass/fail" (actually "mastered/not mastered") as evidence of mastery of the corresponding objective or benchmark.

Pretest for need

A pretest for need tests the terminal objective(s) of the instruction to follow. It can serve three purposes:

- *Placement* in the instruction, to determine that the learner does not already know what is to be taught.
- *Routing* around the instruction, in individualized instruction, by allowing learners to skip instruction on what they already know.
- *Evaluation* of effectiveness of the instruction, where there is a need to demonstrate learning gain. In this type of evaluation, this kind of pretest provides the beginning level of understanding of what is to be taught, and it can detect problems with learners who know too much or too little to be representative of those who are intended to be in the study.

The closer to the beginning of the instruction this test is administered, the better the measurement will be of learning at the beginning of the instruction.

Like the lesson quiz, it is important for these tests to be a highly valid measure of what is taught in the instruction. If the instruction is in turn defined by standards, then the pretest for need will be a valid measure of the corresponding benchmark or objective.

When these tests are used for placement or routing, they are low-stakes, since the only consequence of measurement error would be to waste a small amount of the learner's time (by teaching or skipping a lesson inappropriately). Therefore, reliability can be moderate, and the test cost and test length can be modest.

When pretests of need are used to provide baseline data for evaluation of instruction, it may be appropriate to use a high-stakes test in order to get the superior reliability offered – if one is available which corresponds closely to the content taught. In standards-referenced curricula, this usually rules out use of norm-referenced tests; a state standards test (which is competency based) may be an appropriate measure, if administration is at the right time (such as the beginning of a semester).

These tests usually vary in length, depending on their purpose. Pretests of need used for routing are often very short and are given immediately before each lesson. They can simply be an alternate form of the lesson quiz, but with different questions.

Pretests of need used for placement and evaluation tend to be somewhat longer because they often cover more than one lesson's content. They can simply chain together a number of lesson quizzes (with questions different from those used in the actual lesson quizzes), or they can be specially written for the purpose.

To minimize testing time, for many curricula the PLATO© system offers the FASTRACK tailored testing system as a pretest of need used for placement. It is described more fully in Part 4 of this paper.

Pretest for readiness

A pretest for readiness tests the mastery of prerequisite knowledge of the instruction to follow. It can serve three purposes:

- *Admission* to the instruction, to show that the learner has mastered knowledge and skills assumed by the instruction to come.
- *Diagnosis of deficits* in prerequisites, for assignment to individualized instruction.

- *Evaluation* of instruction, if the evaluation plan requires a check to be sure that learners are ready for the instruction. Typically, this kind of test is done in combination with a pretest of need.

The closer to the beginning of the instruction this test is administered, the better the measurement will be of the learners state at the beginning of the instruction.

All instruction assumes a starting point of some prior knowledge on the part of learners. It is important for pretests of readiness to be a valid measure of what these assumptions are by the instruction to come. If the instruction is in turn defined by standards, then the pretest for need should be a valid measure of the corresponding benchmark(s) or objective(s) which are prerequisite to what is about to be taught. This test content is very different from the pretest of need, discussed above.

When readiness tests are used diagnostically, they are low-stakes, since the only consequence of measurement error would be to waste a small amount of the learner's time (by teaching or skipping a lesson inappropriately). Therefore, reliability can be moderate, and the test cost and test length can be modest.

When pretests of readiness are used for admission, they may be high-stakes (as in college admissions). In this case, test reliability must be high, and the cost and length of development and delivery can justifiably be high. However, if the admissions decision is relatively low-stakes (such as admission to an enriched or advanced placement program), it may be appropriate to use moderate-reliability tests with their shorter length and lower cost.

When pretests of readiness are used for evaluation of instruction, the purpose is to verify that all learners are ready for the instruction to follow. This is an important evaluation consideration, because research such as Bloom's (Bloom 1982) has shown that approximately half of the variation in achievement found at the end of instruction is related to variation in achievement at the beginning of instruction. In most cases, however, this is a low-stakes use and the test can be relatively short and low cost.

Cumulative (course and unit) tests

The purpose of cumulative tests administered at the end of a unit or course is to measure longer-term retention of what was taught, and (if course content builds on itself) to measure the learner's ability to integrate the content of the whole unit or course.

Cumulative tests given at the end of units and courses typically sample the terminal objectives of the lessons in the unit or course, because an exhaustive test would be too long. If the content builds on itself (so that, for example, a learner who has mastered lesson 5 must also have mastered lessons 1-4), then it is appropriate for the cumulative test to concentrate on the higher level (and later) content in the unit or course. If the content does not build on itself, then the cumulative test should randomly sample all the content taught.

Instructors often give more weight to unit and course tests than they do to lesson quizzes and mastery tests, but this may not be appropriate. Often lesson-level tests have many more questions per objective, and taken together they test course content more thoroughly than usually is possible with a cumulative test.

Validity of cumulative tests should be judged by how closely tied they are to the course objectives – which in turn should be tied closely to appropriate standards and benchmarks. The tests should therefore be competency-based and criterion-referenced.

Cumulative tests are generally low-stakes (since the usual consequence of a measurement error is a grade which is too low or too high), so reliability can be modest. This permits test length and cost to be relatively low.

Certification & Standardized Tests

State competency tests and other tests which certify attainment of standards or a defined level of achievement also are given at the end of a block of instruction. They typically are used for high-stakes purposes, such as promotion, graduation, or employment. With accountability now national policy, they also are used to demonstrate quality of the educational system, and the jobs of school administrators and teachers may depend on these tests – a high-stakes use, if ever there was one!

Validity of these tests should be determined by demonstrating how closely they are tied to standards. The tests should be competency-based and criterion-referenced. Norm-referenced tests (also called standardized tests[1]) are defined by a domain specification which is not based on standards, so they should not be used to certify mastery of standards.

Note, however, that practicalities of test length limit these tests to a domain sampling strategy. Thus they are not suitable for detailed instructional management decisions because they do not test every detail of a standard or benchmark. The information they provide is at a level of detail suitable only for general decisions, not detailed diagnostic prescriptions.

Other factors limiting the utility of these tests is how infrequently they are given, lack of detail in reports, and delays in receiving reports – often months after test administration. These factors make the test results nearly useless for making real-time instructional management decisions.

[1] It is fairly common, but technically incorrect, to refer to all high-stakes tests as "standardized." The process of standardization, also called *norming*, applies only to norm-referenced tests.

Note that the importance of these tests has created the need for similar "practice" tests which can be administered on demand and scored immediately with detailed prescriptive reports.

Reliability of certification and standards tests should be high, with the associated requirements for relatively long tests and high costs of development. However, there has been considerable criticism of the validity and reliability of many state standards tests (Marzano and Kendall 1996). Use of low-reliability, low-validity tests for decisions with major consequences has resulted in successful litigation in other circumstances, and the same may happen with the state tests.

Testing and *No Child Left Behind*

The recent Federal *No Child Left Behind Act* (NCLBA) raises the stakes for educators, and makes effective implementation of standards-aligned curricula even more critical. The Federal legislation has established these goals for the states which relate to tests:

- Development of competency-based tests which align to standards
- Near-mandatory participation in the *National Assessment of Educational Progress* (NAEP), with the expectation that state tests will be consistent with the NAEP tests.
- Annual testing at every grade level
- Over a period of years, expansion of standards and tests to include science.
- Disaggregated reporting of progress, by subgroups of students identified by race, ethnicity, handicap, and economic status.

The danger is that educators may treat these new testing requirements as additional "layers" which only take time away from the curriculum, as discussed above. The intent of the law, however, is to drive standards into the everyday practice of every classroom at every grade level.

This makes it even more important for schools to assure that *all* their daily testing practice is aligned to standards and of adequate quality to allow effective data-based decision-making. Since teachers vary widely in their test writing skills, this is a major challenge. Furthermore, aligning the tens of thousands of test items in use in a typical school district to standards is a daunting task.

Automated systems, such as PLATO© *Orion* GATE and *PLATO© Link,* can help. By providing banks of tens of thousands of pre-aligned, carefully reviewed and quality controlled test items, teachers have at their disposal a flexible, customizable and powerful tool to make standards-aligned testing a reality in everyday classroom use.

Automated testing has additional important advantages to teachers and administrators. For teachers, the systems make it possible to administer individualized tests for each learner, thus supporting self-pacing and individualization. In addition, test scoring and record-keeping are entirely automated – a huge time saver. For administrators, the systems provide real-time, daily updates on progress toward standards, for the entire district or broken down by school, grade, classroom, or subgroup of students. This makes it possible to spot areas in need of assistance immediately – thus avoiding nasty surprises when annual testing time comes around.

A further implication concerns implementation of mastery learning. NCLBA's requirement that 100% of students meet appropriate standards places considerable emphasis on programs of instruction which are flexible enough to adapt to the needs of every individual student. Doing so requires teachers to make frequent – even daily – decisions about what each individual student should be doing: literally hundreds of instructional management decisions each day. Furthermore, the decisions must be made based on data which shows what every individual student has learned, and where problems in understanding may exist. Simply collecting this data, marking the papers, and maintaining the needed records is a task of which only superhuman teachers are capable, if it is done manually. This is why paper-based mastery learning systems almost inevitably collapse under their own weight.

Automating testing, marking and reporting makes mastery learning possible. Furthermore, the PLATO courseware makes possible the complete individualization of pacing which is essential to mastery learning. Taken together, these technologies free the teacher to move to a "guide on the side" role in an effective mastery learning system.

Reliability and Validity Considerations for Standards-Referenced Tests

In the standards-referenced world, all tests – pre-, during, and post-instruction -- should be competency-based and criterion-referenced. While we might wish all tests to be perfectly valid and reliable, issues of test cost and length require instead that educators balance the requirement for validity and reliability against the purpose of the test. This is the significance of the distinction between "high stakes" and "low stakes" used above.

Test quality: high-stakes and low-stakes tests

Tests are used to make decisions. If the decision has major consequences, such as admission to a school or educational program, graduation, employment or promotion, then stakes are high: the consequences of an incorrect decision are important and lasting. If the decision has minor consequences, such as placement within a curriculum, control of progress through a curriculum, or cumulative testing for a course grade, then stakes are low: the consequences of an incorrect decision are usually limited to some wasted time. For example, if learners are placed at too low a level in a curriculum, they might be assigned some unneeded study.

The distinction is important for a number of reasons. Recommendations of standard professional practice in testing (American Educational Research Association., American Psychological Association. et al. 1999) suggest that tests of only the highest quality (and cost and length) be used for high-stakes tests, and this recommendation has frequently been supported in civil court cases. Low-stakes tests, on the other hand, are not held to such high standards of quality, cost and length.

Test quality is expressed in terms of validity and reliability. Both concepts guide the interpretation of the data generated by the test, rather than the appearance of the test items themselves. Tests of high quality (high stakes tests) have higher validity and reliability, because the tests have been developed using a great deal of research as well as technical care. This is what makes high quality tests costly to develop, and often relatively lengthy. By contrast, a moderate-quality test (low-stakes test) rarely is developed with a similar level of research, and development cost is considerably lower.

Validity

As discussed above, in the standards-referenced world validity is determined by how well a test corresponds to standards. There is no simple statistical procedure to prove validity:

> There is an implicit *two-step rationale*: First, relevant knowledge and skill important to domain performance are delineated by means, for example, of job or task analysis or other sources of domain theory; second, construct-valid measures of the important knowledge and skill are selected or developed. Test items and tasks are deemed domain relevant because they ar4e presumably construct-valid measures of relevant domain knowledge and skill or, more generally, of relevant domain processes and attributes (Linn, National Council on Measurement in Education. et al. 1993).

Therefore, it is important to establish review procedures by which trained panels of experts compare items to their standards and judge if they correspond in content, Taxonomy level, and difficulty.

Tests also are sometimes shown to have bias by favoring certain genders, ethnicities, economic groups and student populations. These are questions of validity: a perfect test would measure only the knowledge it is designed to, and all test takers who know that knowledge would perform the same way on the test. Precautions during test writing and editing can minimize these sources of validity issues, and test item reviewers should be trained to address these issues.

Sometimes research projects are used to establish statistically the validity of a test. However, these methods for establishing validity generally are appropriate only for high-stakes tests. For example, to guard against bias in high stakes tests it may be worthwhile to compare scores on the test among equivalent student populations of various profiles, in order to detect differences. An advanced statistical technique for scoring tests, called item characteristic curve modeling, is possible only on high quality tests, but has yet to be applied to state standards tests (though the next generation of tests may do so).

The most common way to improve validity of a norm-referenced test is to give it to different groups of students and compare the results to see if the test mirrors the actual differences between the students, as measured by some other means. For a competency-based test, this method is less useful; instead, the best way to improve validity is to see how well the test questions map back to their corresponding standards and benchmarks. This is usually done with expert raters.

Reliability

Reliability is the degree to which a test works in a predictable way. In principle, if you give a test to a particular learner twice, a perfectly reliable test would always produce the same score[2]. Thus, a valid test also has to have good reliability (though it is possible to have a reliable test which is not valid).

Procedures for establishing reliability of competency-based tests differ from those for norm-referenced tests. Well-recognized principles for improving the quality of test items, found in standard measurement texts, apply to tests of both types, and a rigorous review and editorial process should apply to all test items as a means of improving reliability. Statistical techniques, such as item analysis, take on a unique form for competency-based tests, however, and are commonly applied only to "high-stakes" tests. Using item analysis, "high stakes" tests discard weak items (often as many as 9 items will be discarded for every one used). When tests are used to make "high stakes" decisions such as graduation and admission, it is worthwhile to go to considerable expense to assure the reliability and validity of the test. When the purpose is "low stakes," the tests can be shorter and of lower reliability and validity. It is important not to make "high stakes" decisions with "low stakes" tests. However, it is impractical and too costly to make "low stakes" decisions with "high stakes" tests.

[2] In practice, a learner learns a bit just by taking a test, so a re-administration of the exact same test would produce a higher score. To prevent this problem, it's common to use two parallel forms of the same test. This is called test-retest reliability.

Note that it is not true that some question formats are inherently more reliable or valid than others. For example, teachers often mistrust multiple-choice questions – perhaps because they have seen (and written) so many that lack reliability and validity. But other question formats, such as essay questions, can have equally bad problems with reliability and validity unless they are carefully designed and scored. The same applies to performance-based measures (such as projects): the apparent realism of the task (face validity) often masks significant reliability problems in scoring the work products, and the poor reliability in turn limits validity.

Fig. 3, below, summarizes the tradeoffs between test purpose, length and cost, and reliability.

Purpose (Stakes)	Length/Cost	Reliability
Low	Short/Low	Low
Moderate	Long/Low	Moderate
High	Long/High	High

- Fig. 3: Tradeoffs between test purpose, cost, and reliability

Accountability has placed new emphasis on the need for tests which accurately mirror the standards and benchmarks (and thus are valid and competency-based), and which are of appropriate reliability. For high stakes purposes such as state standards tests, reliability (and cost and length) should be high. For low stakes purposes such as placement, progress and end-of-unit tests used in the course of instruction, reliability (and cost and length) should be moderate. Teacher-written tests, because of their often poor reliability and validity, may be unwise even for low-stakes purposes.

Reliability and Validity Procedures Used for PLATO Tests

All PLATO© test systems have been developed using procedures and design standards that follow accepted professional practices for low-stakes competency-based, criterion-referenced tests.

Validity

For competency-based tests, the critical validity issues concern alignment of test items to objectives, benchmarks and standards (content validity), as well as procedures for domain sampling. PLATO Learning makes no claims of predictive validity for any of its tests. Allowing for variations due to requirements for each test type, these general steps are followed in development of all PLATO tests:

1. A detailed content map is developed, by use of task analysis procedures (for courseware) and detailed analysis of state and national curriculum standards and benchmarks (for all products). This analysis is typically much more exhaustive than the domain specification which suffices for norm-referenced tests, and takes into account issues of knowledge type and structure and Taxonomy level as well as topic.

2. A *test specification* is developed, with model items corresponding to each objective to be tested, item specifications such as difficulty level (cognitive load), reading level, and typical errors to capture, as well as number of items to be developed per objective. This is in place of the domain sample specification used in norm-referenced tests.

3. The test specification document is then reviewed by subject matter experts and instructional design specialists. Criteria for the review include content accuracy and distribution of items, item writing style, correspondence to objectives and standards in content and Taxonomy level, and technical feasibility.

4. Where model items have been published to correspond to the target curriculum standards and tests, or where research has demonstrated that a particular item format is preferable, the PLATO© items follow these recommendations. Item formats vary according to the system. Most common are 3- and 4-choice multiple-choice formats. PLATO courseware also includes a variety of constructed-response formats. Binary response (true/false, etc.) formats are avoided unless the reference standard uses them, because of their inherently low reliability.

   Some practice tests also emulate the exact physical format of the target high-stakes test. This is called *full idiom coverage.*

   Also of note is the ability of the National Writing Practice Test to grade responses to essay questions for content. This system was developed in partnership with Educational Testing Service. It uses the *eRater* essay test grading system, a highly valid and reliable technology used for the essay components of many ETS high-stakes tests.

5. Guidelines are also applied in item writing and review to minimize bias by age, gender, economic status, race and ethnicity. For example, care is

taken to use vocabulary, figures of speech and contexts which are widely identified and understood, and to avoid cultural references which are likely to be too specific, such as references to holidays, sports, colloquial and regional expressions, unhealthy foods, gender or ethnic stereotypes, and the like.

6. All items are reviewed for content accuracy by an independent subject matter expert (SME).

7. All items are peer reviewed for clarity and style, as well as sound instructional design.

Reliability

Additional design guidelines and development practices help to assure reliability.

8. All items are peer-reviewed and edited for use of recommended item writing practices appropriate to that item type. Guidelines used are drawn from standard psychometric references, such as Gronlund (1993), Haladyna (1997), and Osterlind (1998).

9. Item formats are designed to minimize extraneous difficulty in item comprehension and answer entry due to limitations of the user interface. Thus, for example, keyboarding skills needed are minimal for most item formats, and care is taken to keep all information needed on screen at once, wherever possible.

10. All items undergo technical review and testing to verify that screen displays, user interface and answer analysis work correctly.

11. For pre-defined tests, additional quality controls are applied in a final review. Each test is taken with all questions answered incorrectly. The process is repeated with all questions answered correctly. Each question is given an in-depth check, including evaluation of formatting and appearance. The following questions are asked:

    - Are there enough questions to test for mastery of a skill?

    - Are the test items age- and grade-level appropriate?

    - Does each skill appear in the listed objectives and does each objective appear on the test?

    - Is each question answerable (not vague, subjective, or incomplete)?

> The test is taken again, answering the questions correctly. Each possible answer is checked in-depth. A time-on-task analysis is performed on the test, to determine whether a student could take it in less than 45 minutes. The following questions are asked:
>
> - Is there a correct answer for each question?
> - Do any answers appear twice?
>
> Each test is generated repeatedly to check all possible combinations of questions and answers. Any tests that fail to pass all criteria are edited and put through the same rigorous process again.

To minimize item exposure in re-testing, all testing systems use random selection of items from a defined pool of items (exceptions are certain reading comprehension tests, which randomly select passages accompanied by clusters of items, and certain practice tests which are fixed format). The *PLATO© Link* system adds a most-recent use algorithm to further minimize item exposure. Predefined tests select between 3-5 items per objective; user-defined tests may select any number of items per objective. Refer to the next sections for specifications of item pool size by test.

At this time, statistical analysis of item reliability is performed only on the *PLATO© Link* item bank. Test reliability statistics are not available.

Part 3 of this paper summarizes and compares the two general-purpose testing system options available to PLATO users: PLATO© *Orion* GATE, and *PLATO© Link*. Part 4 summarizes and compares the testing systems which are embedded in PLATO courseware and designed to support its use.

Part

3

PLATO Learning Comprehensive Testing Systems

PLATO Learning offers two comprehensive testing systems for reading, writing and mathematics curricula. These systems provide a convenient way to select or construct standards-referenced low-stakes tests, to mark them and to report results. They are linked to instructional resources (including PLATO courseware) by reference. The Grading and Testing Engine (GATE) is part of the *Orion* curriculum planning and management system, and is an ideal tool for *Orion* users. For clients who do not use the *Orion* system, or who need a larger item pool, PLATO Learning offers *PLATO© LINK.*

PLATO Learning also offers the Simulated Test System (STS): a series of practice tests which emulate particular high-stakes tests in content and item format and which provide prescriptions directly to PLATO courseware. The K-12 simulated tests are delivered through the *PLATO© LINK* system. Other simulated tests (of GED 2002, Pre-Professional Skills Test, and the National Writing Test) do not require a *PLATO© LINK* subscription and do not provide access to other *PLATO© LINK* facilities.

The National Writing Test was developed in partnership with Educational Testing Service. It uses their highly-regarded *eRater* essay question scoring technology – the same technology used to score essays in ETS' high-stakes tests. The National Writing Test provides essay questions typical of those included on state competency tests. The full text of each learner's essay is scored for content as well as mechanics. The learner instantly receives scoring information and model essays for comparison.

Both PLATO© *Orion* GATE and *PLATO© LINK* address a wide variety of key needs for internal, low-stakes testing, including:

- **Transparency**: The systems are understandable and clear to students, parents, and educators. They provide teachers and students with timely, effective feedback to facilitate progress toward meeting standards, and ensure that teachers of all grade levels work as a team to meet standards responsibly and responsively. Parents receive user-friendly and helpful information about student performance.

- **Practicality:** Tests are easy to compile and administer, and students and teachers find the system easy to use. The tests are minimally intrusive, since they can provide frequent, short tests rather than time-consuming, one-time tests, and they provides unlimited access for all parties. They are closely aligned to regular classroom instruction and provide a continuous improvement model with a data-driven decision-making vehicle.

- **Capabilities:** Both systems can be used to provide formative testing, online prescriptive planning, benchmarking, and data-driven instructional decisions for instructional grouping, classroom-level re-teaching, individualized remediation instruction through PLATO courseware, and individualized academic assistance planning.**Timely, Flexible Reporting:** Unlike remotely-graded paper-and-pencil tests, both systems' online tests are instantly graded. Results are reported in a wide variety of formats which support demographic disaggregation and data-based decision-making in real time.**Assessment types:** Both systems provide online assessments suitable for grade level assessments, strand tests, and skills subtests. Both systems can be used for practice on content for standardized tests, but PLATO© LINK has the additional capabilities of the Simulated Test System.**Customization.** In addition to pre-defined tests, teachers can easily select items for their own tests. GATE also provides the option of locally developed tests in a variety of item formats, including performance-based assessment capability. **Tied to PLATO© Instruction.** Both systems provide prescriptive links to PLATO courseware[3], for highly effective, targeted instruction and remediation in self-paced, on-demand environments.Both testing systems are available only through Internet subscription.[4] They are browser-based, and thus will run on PC or MacIntosh with Internet access.[5]

One of the essential features of both systems is their ability to provide educators with diagnostic reporting on individual students, classes, schools, and districts. Reports are Web-based and generated in real-time. Therefore, educators are able to access and view these reports twenty-four hours a day, seven days a week, immediately after test administration. This is a capability not available from any state or standardized test. In addition, both systems allow administrators to view cumulative or comparative reports at the district, school, and classroom level, and both systems support disaggregated reporting by demographic group. And, both systems provide secure access for parents to see information relevant only to their child, if administrators wish.

Note that tests may be administered on-line or via paper and pencil. Online administration allows automatic scoring and reporting. Online, if desired, each

[3] GATE will be linked to PLATO Web Learning Network courseware in Fall, 2002; PLATO LINK has the capability now.

[4] PLATO's other standards-referenced curriculum planning system, *Polaris*, is Internet based but is less costly to implement in a client-hosted Intranet configuration. However, it currently has relatively small item banks and generates only paper-and-pencil tests. It will not be discussed further in this paper.

[5] For current minimum platform specifications, refer to product data.

learner's test can be a unique sample from the system's item banks, so it is possible for each test to be different. This makes it possible for individualized programs to administer tests at different times, on demand, and still maintain test security. Both systems include extensive item banks. In addition, GATE has a system for instructors to enter their own test items, in a wide range of formats.

In both systems, the tests are linear, and the learner is always presented with all the questions on the test; there is no early termination of testing for learners who are doing particularly well or particularly badly. The table on the next pages compares and contrasts the two systems and provides additional details of functionality.

PLATO Comprehensive Testing Systems for Reading, Writing and Mathematics

	PLATO© Orion GATE ver. 4.0	PLATO© LINK User-Defined Tests	Simulated Tests
1. *What does it test?*	Math (basic through Trig.) Language Arts (reading, writing, study skills) Referenced to standards	Math, Reading, Language Arts Referenced to standards or textbook scope & sequence	Many state standards tests and standardized tests in math, reading, writing, GED II & Pre-Professional Skills Test (PPST)
2. *What item formats?*	MCQ[6], 4-choice, single answer Teacher-provided items in other formats Order of alternative choices shuffled at test administration for each student.	MCQ, 3-5 choice, single answer	Same as format & idiom of corresponding standardized test.[7] Writing: essay questions with full response recorded & scored for content and mechanics
3. *How many items per objective in item pool?*	Language Arts: 3-5/objective Math: 10/objective Total Pool >50,000 items[8]	30 to >100 Total Pool >130, 000 items	Number of items is based on the actual test being simulated.
4. *How many items per objective in a test?*	3-4 items, randomly selected	Teacher-specified	Follows original test
5. *What levels/grades?*	Math: 3-12 Language Arts: 2-12	3 – HS exit exams	Follow original test (often 3 or 4, 6 or 8, exit)
6. *What test structure?*	Linear	Linear	Linear
7. *Early termination?*	No	No	No
8. *Prescription*	% right per objective Resource recommendations + PLATO courseware [9] Best practices	% per objective Resource recommendations & PWLN[10]/Pathways[11] IDP[12] generated	% right per objective Resource recommendations & PWLN/Pathways IDP generated

[6] MCQ = multiple-choice questions

[7] Test format varies by test. Some emulate content and test format, while others emulate content only. See product descriptions for details.

[8] 15,000 items included in current version

[9] Fall 2002

[10] PWLN=PLATO Web Learning Network (Internet delivery)

	PLATO© Orion GATE ver. 4.0	PLATO© LINK User-Defined Tests	Simulated Tests
9. *How is start & end point set?*	All students take full test	All students take full test	All students take full test
10. *Student Feedback*	Immediate feedback on individual items right/wrong with prescriptions for additional study resources + PLATO courseware	Immediate feedback on individual items right/wrong with prescriptions for additional study resources + PLATO courseware	Score + Automatic prescription to PLATO courseware
11. *Reporting*	% right/objective, Total % right, Class, School, District summaries Data export for additional analysis/data mining.	Automatic prescription to PLATO (for full length tests only) % right/objective, Total % right, Class, School, District summaries	Automatic prescription to PLATO % right per objective Total % right Individual, Class, School, District summaries
12. *Look & feel*	Browser: text + line graphics or .PDF (printable for hand scoring)	Browser: text + line graphics, or .PDF w/ bubble sheet (printable for hand scoring)	Browser-text + line graphics
13. *Platform*	Browser with Orion server. Works with Windows, Mac, PDA's browsers.	Browser with PLATO© LINK server	Browser with PWLN + PLATO© LINK server
14. *What alignment used?*	State standards (free) or any accountability document (for fee), based on PLATO© alignment standards.	State standards, district standards (user-supplied), tests, textbooks. May align to PLATO curricula. Based on external alignment standards, indexed to PLATO alignment data base.	Follows state standards or standardized test (domain specification may be based on standards)
15. *How are items selected?*	User selects items based on lowest level of state/local standards	User selects items based on state standards, tests, or textbooks.	Pre-defined, based on state standards tests.
16. *Customization Options*	Teacher – supplied items (many formats) Teacher-specified tests (user-defined)	Teacher-specified tests (user-defined)	None
17. *Purpose*	• Pretest (for evaluation) • Quiz • Unit test • End of Year Test	• Pretest (for evaluation) • Quiz • Unit test (by standard or textbook) • Cumulative test (semester/year)	Practice tests to prepare for high-stakes tests

[11] Pathways = LAN delivery

[12] IDP=Individual development plan. Works by exempting learner from assigned modules which the test shows are not needed, so only the needed modules remain in the assignment.

□Fig. 4: Comparison of PLATO Comprehensive Testing Options

PLATO *Orion* Grading and Testing Engine (GATE)

PLATO *Orion* GATE is designed to simplify grading and assessment. It can be configured to diagnose student performance against state objectives, provide remedial prescriptions for student weaknesses, and (with PLATO courseware) instruct students based on their identified needs. Individual academic assistance plans can be generated for each student. GATE also has flexible scheduling and scoring options, and provides detailed analysis and reports.

An Overview of GATE

GATE is a completely integrated online test engine that uses a proprietary database to align standards and benchmarks to a vast array of on-line -- and on-site -- resources for classroom instructional purposes. At the same time GATE also provides teachers with accurate data from which to analyze student performance and make informed decisions regarding corrective action. Included in GATE are individualized Academic Assistance Plans aligned to each state's standards and benchmarks. These provide student-unique feedback, with recommended re-teaching resources for any objective for which the student has not demonstrated proficiency.

Using GATE, teachers can create and administer tests, quizzes and questionnaires on any math or language arts strand. More importantly, these assessments can be matched to specific state standards and benchmarks and may be included in lesson plans or individual education plans (IEPs). GATE will grade and analyze responses to questions in multiple-choice test formats. Teachers can choose from a bank of thousands of items, and may enter their own items in a wide variety of formats.

Any changes in standards or benchmarks are quickly reflected in the PLATO© standards database, and the database is adjusted accordingly. As a consequence, any resources or test items associated with those standards or benchmarks are automatically re-aligned. This applies in the case of a change, deletion, or addition of new standards, as well as for any new resources or tests.

GATE contains an editor function that allows the PLATO Learning's in-house team of experienced alignment and test specialists, all experienced teachers, to enter items and create tests.

The GATE item bank has been licensed from a third-party vendor, a recognized provider of tests and test banks to publishers and schools across the U.S.

While the tests are designed to be taken online, there are no restrictions with regard to printing tests and administering them in paper-and-pencil form. As soon as a student finishes an online test, it is automatically scored, with feedback and a remediation plan provided immediately. Scoring occurs by means of a complex series of algorithms, with the results matched to specific curriculum objectives from which remedial action can be taken. If tests are administered via paper and pencil, the ability to have them automatically scored and reported on by GATE is not available. However, teachers may print a scoring template for hand-scoring.

Reports may be generated for each student, class, or group of students, and may also be accessed by students and parents if desired. While teachers can see their students and class, and principals can see all or part of their school. Administrators can see cumulative data for their district, individual schools, classes and students, and can disaggregate data for reporting by demographic group.

GATE is part of the PLATO *Orion* system. Note that many functions needed for a standards-based implementation are in the companion Planner module of *Orion.*

All access to GATE is by unique IDs and passwords generated and administered by a locally-appointed Site Manager. The GATE system is implemented in the Oracle database, which is noted for its security, and is further protected by robust firewalls and a secure data center.

Alignment to Standards

When test items, web site URL's, textbook chapters or other resources are entered, they are "tagged," or identified, as related to certain standards or skills, by using a standard thesaurus of keywords as well as content names. This allows them to be correlated to the state standards and the most detailed benchmarks available, more effectively than is possible with a simple text search. Specialists also code items by grade level and (for test items) by Taxonomy level.

Thus, both for preconfigured tests and when teachers create their own tests from the GATE item banks, they are associated with specific benchmarks, and may be further specified by grade level and Taxonomy level. However, if teachers choose to enter their own items, this alignment is not possible.

Tools available in GATE

GATE includes the following tools:

- TestManager
- TestScheduler
- Feedback Report Generator
- Question Browser
- Quizmaker

- Gradebook
- Academic Assistance Planner
- Diagnostic and Achievement Tests
- Media Center
- Accountability Reporting

Refer to product descriptions for current details of these tools.

Reports

An exceptionally wide variety of reports are available in GATE. The basic teacher's report lists test(s) and a synopsis of percent scores, with additional reports available (see below). A display of all students who have completed a scheduled test, including their score, has icons and links to additional reports. These reports include a Question Report, Curriculum Report, Feedback on the student's actual responses, and a Remediation Report.

- The Question Report shows the number of responses and correct responses for each question with an icon to indicate how each student answered any particular question.
- The Curriculum Report provides valuable information for all students who took this test that shows percentage scores for each standard and benchmark measured by this test. A similar report can also be obtained for each individual student.
- To obtain feedback, a teacher or student can click on the Feedback icon to see a copy of the student's test with student responses and correct responses indicated..
- Remediation Reports can be generated for all students or each individual student. This report lists websites, textbooks, software, or other resources that can be of help in re-teaching the skill or concept. This can become the basis of a personalized academic improvement plan for each student or class.

User- configured reports can be generated on demand. Configurable data available includes:

- Results by student
- Results by class or class period

- Results by campus/district
- Results by benchmark
- Item analysis: correct and incorrect responses by student, class or period, campus, district
- Item analysis: correct and incorrect responses by Student Expectation not mastered
- Item analysis: correct and incorrect responses by item.

A **data disaggregation** capability by ethnic, race and socio-economic status variables has been added to the current data base.

Additional reports may be provided at the discretion of the school or district to the student and (custodial) parent of each student. These reports may be printed (locally or on a network), or privileges granted for online access of these reports.

GATE diagnostic tools include capabilities such as: individualized Academic Assistance Planning, item analysis with item-based prescriptions, and dynamic information on student, class, and school performance against benchmarks and standards.

Item Pool

The GATE test item pool consists of developmentally appropriate multiple-choice items with at least four response choices. Many of these items contain associated graphics and/or paragraphs for problem solving and critical thinking. Distracters for the multiple-choice items reflect both a wide range of typical student errors and "oddball" distracters where appropriate.

The items in the pool are multiple-choice, but GATE supports a variety of test item formats. Teachers can create their own online assessments using these formats or they can use the items from the database. Test item types that are supported include:

- Multiple-Choice
- Short Answer
- True-False
- Essay
- Project or Performance Assessment

Math Item Generation Algorithms

An additional strength of the GATE math item bank is that it is constructed with the use of over 3,000 item algorithms. Currently aligned items within the database can at any time be easily enhanced by PLATO Learning specialists through the generation of additional assessment items based on each item's unique algorithm. This ensures that there will be sufficient items for secure questions as well as those items that are available for use in teacher-created assessments.

In the current release of *Orion*, about 15,000 items have been correlated and are available. An additional 35,000 items, plus additional item algorithms, remain to be correlated and made available in future releases of the system. Thus, it is possible to quickly respond if a particular gap in item coverage emerges.

Validity and Reliability

Test items are reviewed for content accuracy, correspondence to standards, freedom from bias, coverage of standards, and sound item writing style using the procedures summarized in Part 2 of this paper.

Preconfigured Tests

GATE includes preconfigured tests for most grade levels and strands in many states (refer to product data for a current list). These are cumulative tests which may be used as pretests of readiness or need at the beginning of the school year in order to measure the preparedness of each student, or at the end of the year to measure progress.

Teacher-defined Tests

The item pool's coding by standard, grade level and taxonomy gives teachers flexible and powerful options to configure their own assessments at any point during the school year, using a simple web interface. Items can be identified via the proprietary search engine used by the PLATO© database and accessed through GATE.

PLATO© LINK System and Simulated Test System

The *PLATO© LINK* system provides preconfigured and teacher-specified tests in reading, language arts and math which may be specified by reference to state standards and benchmarks, or by reference to textbooks in use. Supporting each standard is a collection of resources including pre-screened Web links (URL's) and sample lesson plans. *PLATO© LINK* also incorporates the *Simulated Test System* (STS), which provides practice tests corresponding to many state and national high-stakes tests in writing, reading and math. Powerful reporting capabilities are available for administrators, teachers, students, and parents.

A unique feature of the *PLATO© LINK* system is that it is mapped to the scope and sequence of the math, reading and language arts textbooks (or any basal scope and sequence) chosen by any client. Thus, at implementation, the site is already set up to match the textbooks being used by all the students in the building or district.

Part of *PLATO© LINK* is the Simulated Test System (STS), which provides practice for a variety of high-stakes tests. Included are many state tests, and STS has the major norm referenced tests such as the SAT-9 exam, Terra Nova, and ITBS.[13] In addition, STS includes the National Writing Test, which consists of essay questions of the types typically found on high-stakes tests, with automated scoring for content as well as mechanics.

PLATO Learning's partner in *PLATO© LINK* is The Princeton Review, a national leader in testing and test preparation.

PLATO© LINK has been designed to improve student achievement through these capabilities:

- **Individual/Class assessment.** Enables educators to assess, analyze, and monitor individual student and overall class academic performance. Students can even generate their own practice tests.

- **Customized content**: Tailored by state standards and benchmarks, by state or national test, and by grade; aligned to textbooks being used in classes (or any scope and sequence)

    - Full-length Simulated Practice State Assessments

    - Administrator Benchmark Testing

- **Individualized prescription**: Identifies students' skill strengths and weaknesses in real time; facilitates data-driven instruction and informs classroom instruction on skill gaps – aligned to state assessments and standards

- **Targeted educational resources**: Exhaustive supply of standards-referenced student/teacher/parent resources including lesson plans, teaching tips, and web links.

Alignment to Standards

PLATO© LINK includes the standards and benchmarks as they have been written for all states, national standards, and specific large districts. Regardless of how

[13] Simulated tests are based solely on publicly released information, so the security of published tests is not compromised. Thus, the simulated tests are consistent with the regulations enforced by the state of California regarding SAT-9 preparation, Virginia regarding the SOL, and other state regulations and publisher requirements.

broad or specific a set of standards is for any given state, alignment specialists have been able to break these standards down into their component parts and thus map them precisely to the relevant math, reading and language arts content within PLATO© LINK. Overall, the content of PLATO© LINK is wholly customized – or customizable (in the event of new math/reading/language arts textbook adoptions or change in scope and sequence or a change in state and tests standards).

Moreover, PLATO© LINK is mapped specifically to the scope and sequence of the reading, language arts, and mathematics textbooks as they map to the state standards and benchmarks. As such, PLATO© LINK presents teachers with the option of creating full-length tests or quizzes which are consistent and reflective of the chapters being addressed in their in-class textbooks. PLATO© LINK is the only on-line educational resource that supports this capability.

Tools available in *PLATO© LINK*

PLATO© LINK provides administrators and teachers with customized views and functionality based on their roles within the educational hierarchy. Specifically, teachers have the ability to use PLATO© LINK as either a day-to-day testing tool or as a periodic benchmarking tool within their individual classes for the purpose of pinpointing both class and individual student strengths and weaknesses. At the same time, both district-and school-level administrators typically utilize PLATO© LINK 's high-level reporting tools to get a cohesive understanding of standards mastery across classes, individual grades, and schools within a district. Administrators also have the ability to make use of PLATO© LINK 's test-creation tools for the purpose of creating benchmark tests for designated schools and students within their districts. As a result PLATO© LINK provides multiple ways of creating and administering assessments and reporting on student, class, and school performance.

Major functions include:

- Test-Creation Tools:
- Pending Assignments:
- Student Progress Reports:
- Remove Students" text link.
- Chapter Selection:
- Alignment to State or Test Standards:
- Test Length/Test Administration:

- Test Creation (Step 2): Skill/Number of Questions:
- Additional Assignment Parameters:
- Test Administration and Scoring
- Teacher and student home pages
- Test Tunnel Page

Further details and sample screens are available in product data.

In addition, PLATO© LINK's Simulated Test System (STS) automatically provides individualized prescriptions to PLATO courseware based on an individual learner's test performance. The system works by exempting learners from a requirement to study those courseware modules which are aligned to the test. The learner then studies those aligned modules which remain to be studied, together with any assigned modules which are not part of the test's alignment but are part of an assigned curriculum structure.

Offline Test Administration

If a test is administered on paper, students have the ability to access a customized scoring interface by clicking on the name of the offline assignment from the Student Home Page. Students will then be able to key in their answers in an efficient manner, enabling them to use PLATO© LINK's scoring and reporting tools.

The Princeton Review created this solution for schools with limited computer access for students. Tests are created to be taken on paper, and students are then able to enter their results in a fraction of the time it would otherwise take to complete a test online.

At the same time, by feeding student test results back into the system, educators and students alike are able to take advantage of key PLATO© LINK functionality, such as its scoring and reporting tools.

Reports

Powerful reporting meets the needs of administrators, teachers, individual students and (if desired) parents.

- Standards-referenced Teacher Reporting
- Administrator Reporting, with support for disaggregation by demographics as required by the *No Child Left Behind Act* (NCLBA).

- Special, secure reports for students and parents

Teachers are able to assign tests or quizzes that are to be completed by a given due date. Through the real-time scoring and reporting capabilities of the on-line utility, teachers are able to quickly determine which students have completed a given assignment. Students are clearly informed of assignments that are overdue, along with the due date(s) of all outstanding assignments. From an administrative perspective, school administrators can assess the performance levels of a given classroom or the entire school, thereby effectively and consistently addressing the need for teacher accountability. Through these monitoring capabilities, PLATO© LINK is fully capable of supporting all aspects of teacher and student accountability.

Teachers are provided with rolling views of the progress of their individual students, as well as their entire student load. The teacher can assess the progress of an entire class or an individual student over time. In effect, a teacher can comparatively review how a student performs over a progression of assignments, in effect improving upon traditional forms of reporting, assessment, and remediation. A student's performance on all generated assignments will be saved and can be accessed from the teacher's class homepage.

PLATO© LINK's new Benchmark Testing System can be used for scheduled administrator-developed monthly benchmarks testing, grade-wide Objective specific quizzes, general assessing needs, all supported with new NCLB-compliant Administrator and Benchmark reporting systems.

Teacher Reports

Teachers have the ability to view both individual student and class-based reports detailing overall scores, and performance based on each skill tested. Relevant information includes each student's name, date he or she completed the test, overall performance as an indication of percent of items answered correctly, and a PLATO© LINK to details about student's performance on each skill on the test displayed.

By clicking on the "Details" icon for a specific student while viewing overall class progress (within the Class Report page), teachers are able to view an individual student's performance on the test being reviewed. Performance is displayed both in terms of performance on individual skills and performance on the actual test items administered. Information on this page includes a list of tested skills, number of questions taken, and the overall performance of the student.

- The Skill column identifies the individual skill being tested within the assignment. By clicking on the name of the individual skill, users have access to instructional resources related to it. Instructional resources are aligned to the state or tested standards, which were used to create the assignment being viewed.

- Teachers can also scroll down to view a student's performance on the test itself. The assignment is displayed with the student's answer, the correct answer (if different), and "Explain" buttons next to each answer choice that provide users with a brief explanation of the correct and incorrect responses.

As a component of PLATO© LINK reporting, teachers are able to review in detail each assigned test by viewing class results with a breakdown of each student's response to individual test questions showing how responses relate to the final scores for individual students and the class as a whole.

Teachers have the ability to click on individual question numbers to view the actual test items administered. The columns also contain each student's answer to individual questions, the correct answer choices (at the bottom of the columns), and the percentages of correct answers per question.

Teachers also have the ability to view diagnostic reports based on a collective progress on assignments taken by the class as a whole. Diagnostic reports display the following information:

- Skill: Provides information on the name of each individual skill within the course chapter selected.

- Tested Skill: Indicates whether a skill appears on the state test.

- Score: Indicates the average score of all students for each tested skill, listed as a percentage score.

- Status: Indicates the number of students tested on the individual skill being viewed.

- Details: Provides a link to a report that shows how each student tested scored on the skill in question.

Administrator Reports

PLATO© LINK not only provides teachers with a powerful pre- and post-test tool within their classrooms, but also serves to provide district and school administrators with additional scheduled benchmark test creation and assignment functionalities. Along with administrator controlled Benchmark Testing, there is a Benchmark Reporting system and Administrator Reporting System where tests, teachers, classes, schools, and grades can be compared within a desired grade and subject, and disaggregated by demographic group as required by Federal *No Child Left Behind* regulations. PLATO© LINK can be used to identify potential "holes" in the curriculum, scope and sequence or text book and be used to highlight areas where teacher professional development could be targeted.

Administrators can view the progress of the entire school or look across several schools with the ability to pull in NCLB-required student demographic data needed for disaggregated reporting. Teachers can view data only for their assigned students. Parents can access the performance reports and records of their child only. Students can only access their individual performance records.

Student Reports

Among the features of Progress Reports, is the ability to identify and present students with a skill-by-skill breakdown of their performance. Once a test is completed, students are able to view their results via an assignment-specific Progress Report. Progress Reports are created when tests are completed, and are designed to offer a breakdown of the skills tested along with each individual question taken as part of the test, and links to explanations of the correct answers.

Students who click on a Skill Helper link for an individual skill within a Progress Report, will be directed to a "Skill Helper" page containing learning resources drawn from the extensive *PLATO© LINK* data base of aligned, screened web sites, textbook chapters, and other resources.

Parent Accounts

Parents can view their children's performance results, and access skill specific-resources using a personalized PLATO© LINK parent account. Personalized usernames and passwords allow parents to track their child's performance throughout the school year and access skill-specific resources and activities written especially for parents to use at home.

After clicking on the name of a specific assignment within the Parent Home Page, parents can access a report that specifies the name of each skill tested, whether individual skills will appear on their child's state test, the number of questions tested per skill, and, test results (both on overall test, and organized by individual skills).

Parents can scroll down to view the actual test taken (including responses selected by the child) along with brief explanations related to each answer choice.

By clicking on the name of an individual skill listed, parents can access pages containing Princeton Review instructional resources and activities consisting of skill-specific exercises to be used by parents in helping their child improve individual skills.

Finally, parents can link directly to activities they can do at home with ordinary household items to help their child improve in applied skills.

Item Pool

There are an ever-growing number of items in the PLATO© LINK item bank database: currently about 130,000. Each item was born of an actual specimen test item provided by test authors, or as a result of The Princeton Review's careful analysis of test blueprints.

In addition to being able to offer a larger volume of questions than most, PLATO© LINK is able to offer a more diverse set of questions, reflecting the many different questions used throughout the country.

Validity

PLATO© LINK uses stringent procedures to maintain content and face validity of its tests. Test items are reviewed for content accuracy, correspondence to standards, freedom from bias, and coverage of standards, using the procedures summarized in Part 2 of this paper.

For simulated tests in *STS*, PLATO Learning and The Princeton Review apply their expertise in tests and item creation by thoroughly investigating the substance and format of the many state tests and the commercially available norm-referenced tests. Source items, taken from non-secure specimens published by the test authors, are used to create a set of questions covering the same content and concepts mirroring the format and representing a range of difficulty similar to, but around, the source item. These items have been created by the thousands for each test the authors have studied. The sheer scale required internal and external resources. The external resources have included some of the same educational publishers who prepare questions for actual exams.

PLATO© LINK's test items are of two different classes:

1. *Full Idiom Coverage* includes items written in the format of the corresponding specimen test, originally for *STS*. These items are available in the dynamic *Create A Test* function of PLATO©LINK, and in two predefined simulated tests per tested grade and subject. Refer to product specifications for details of tests currently available with full idiom coverage.

2. *Full Coverage* includes items which PLATO© LINK has aligned the standards (by grade and benchmark) of a specific state by reference to its standards, benchmarks and tests. Items correspond in content, but not format. These items are available only through PLATO© LINK 's *Create A Test* function. With the addition of the Benchmark System, the predefined tests associated with classes set up this way can be any combination of any state's assessments and/or norm referenced national tests like the Terra Nova, ITBS, or Stanford 9.

PLATO Learning and The Princeton Review recognize that many states have no performance objectives for 'off level' grades. As a result, we made the decision to replicate the objectives of a higher grade to the grades below it. Specifically, in Arizona for example, performance standards of grade 5 are applied to both grades 4 and 5; standards of grade 8 are applied to grades 6, 7, and 8; and, standards of grade 10 are applied to grades 9 and 10. We consider this to be a better method than trying to imagine how schools might sequence learning across multiple grades. Moreover, we consider this to be a sound method because students are working towards the goals of the higher grade. However, our system will be updated as needed to reflect modifications in state standards and/or alternate sequencing requested by local districts, particularly as annual testing under NCLBA becomes a reality.

The process for linking items to standards is very labor intensive, one that we do not leave to technology. Specifically, the Princeton Review & PLATO© staffs have analyzed all state standards and compared these standards to a comprehensive taxonomy of skills in our system. The Princeton Review staffers then map each standard statement to one or more of these skills using proprietary tools created for internal use. Through our system, we are able to correlate all state standards to the skills of the taxonomy. The methodology for correlating questions and resources to standards relies on aligning such content to the same set of skills. Each question is tied to a single skill. Each skill is then correlated to one or more appropriate state standards identified. This methodology allows for a seamless integration between all components within the PLATO© LINK system. Since all items are correlated to the PLATO© LINK taxonomy by subject and grade, the items are correlated to the state standards as well. By maintaining a manual system, we believe that our database is more precise than those that rely on "text string" matches.

The integrity and detail of the alignment process plays a major role in determining the quality of tests drawn from the item banks. In PLATO© LINK, each item is written to address a specific tested benchmark on a specific state or national test. All items are also written to simulate an actual test item and are written in the idiom of that test. Each item is then aligned to one unique objective in the PLATO© LINK data base, and each item is also assigned a grade level. Objectives in the PLATO© LINK data base are derived from analysis of all state and national standards. Thus, alignment to standards is accomplished by relating objectives in the PLATO© LINK data base to corresponding standards.

When constructing a test, the PLATO© LINK testing engine first looks for items in the pool written specifically for the test around which the teacher organized the assessment. If PLATO© LINK does not have items written in the idiom and format of that test, then the search expands to the entire data base of over 130,000 items. The test engine then will pull those items from the idioms of other state and national tests which correspond to the same objective. Thus, a given item can reference corresponding standards from many states, even though it is indexed to only one objective in the PLATO© LINK data base.

Reliability

To assure reliability, items are reviewed and edited using the procedures summarized in Part 2 of this paper. Items are not initially tried out; instead, they immediately become available to users. However, PLATO© LINK does not rely solely on the connection to the source items as a way to guarantee a valid and reliable item pool. Item authors evaluate the items using traditional item statistics. Items are modified and/or removed if they are determined to be too difficult or if a single wrong distracter is unduly impacting student performance.

Depth and Breadth of Items

For each grade there are roughly one hundred math skills and twenty-five reading skills, and a corresponding number of language arts skills. Within each grade there are in excess of eight thousand math questions and two thousand reading questions.

The materials are appropriate for students in grades 3 through high school. At this time, we do not offer content in grades K-2 because of the challenges associated with delivering age-appropriate assessments at this level. Exams at this level often include oral components. We are unable to support and guarantee the performance of audio for testing situations. Similarly, our content at the upper grades covers skills identified by the state as important to a core curriculum, but may not include material for all upper-level content.

Each year The Princeton Review and PLATO Learning author thousands of additional questions, adding to both our breadth and depth of content. For example, language arts materials was included in the system in September 2002.

At this time, PLATO© LINK is a closed system and does not accept new questions from users. This allows us to attend to the quality of each item that is written, evaluating it for its instructional relevance, content validity, and freedom from stylistic errors. This tight control is necessary, as items within our system are available to virtually all of our users. Furthermore, with few exceptions, the depth of our items makes it difficult to justify the creation of additional items within a domain.

Part

4

PLATO Courseware Testing Systems

To support its self-instructional courseware, the PLATO system includes placement tests (of need), module mastery tests for progress control, and cumulative Course Level Assessments (CLA). These tests are available for most secondary/adult PLATO© curricula in reading, writing, mathematics and workplace skills. PLATO© elementary curricula in mathematics include module mastery tests. PLATO elementary reading and math curricula at the K-3 level includes a reading and math inventory testing system, which can be used as a pre-test or cumulative test.

Most PLATO© tutorial modules include one (or more) mastery test lessons with a bank of test items which are carefully aligned to the terminal objective(s) of that module; the module objectives are carefully aligned to state and national curriculum standards. In most cases, the item banks are at least three times the length of the module mastery test. In secondary/adult curricula, these item banks are used to generate pre-defined tests for placement using the FASTRACK system, for progress control in the module mastery tests, and for cumulative testing in the Course Level Assessment (CLA) system. In elementary curricula, placement tests (the Skills Inventory) use their own item pools.

The table on the next pages summarizes and compares the characteristics of these tests.

PLATO Courseware Testing Systems

All PLATO tutorial curricula in reading, writing, mathematics, and work skills[14] have module mastery tests, which are derived from an extensive analysis of state and national standards. Fastrack and CLA are available for the curricula listed.

	Fastrack	Skills Inventory	Course Level Assessment	Module Mastery Tests
1. *What does it test?*	Terminal objectives from each secondary/adult module mastery test in the core products of reading, writing, math[15]	Elementary courseware: *Beginning Reading for the Real World (230 objectives), Math Expeditions* (1400 objectives)	Secondary writing, math	PLATO tutorial module terminal objective(s) (1 test per module, except in some math curricula)
2. *What item formats?*	MCQ, 3-4 choice, single answer	MCQ, up to 5 choice, single answer	MCQ, 3-5 choice, single answer; constructed response	MCQ, 3-4 choice, single answer; various free-response formats
3. *How many items per objective in item pool?*	4 selected from module mastery test item pools[16]	10 unique to test	15-30 (same pools as module mastery tests)	10-60
4. *How many items per objective in a test?*	Up to 4 (reading comprehension: 5) per testlet	5 randomly selected per testlet	Published tests: 4 randomly selected. User-defined tests: user-defined	5-10 randomly selected
5. *What levels/grades?*	Math & Writing: A-I Reading Skills & Strategies: B,C,D,F, I Reading Comprehension: B-M	K-3	None; corresponds to modules	None; corresponds to modules
6. *What test structure?*	Tailored testing by testlet, with "up 1, down 2" algorithm	Linear; follows courseware sequence	Linear, in testlets, 1 per module mastery test	Linear
7. *Early termination?*	Yes. Pass = 4 right, Fail = 2 wrong	Yes. Pass=4 right, Fail=2 wrong	No. Default Pass=80%, may be set by instructor	Yes. Pass = 80%

[14] Selected science, social studies and ESL products also have tests, but these general standards do not apply to them.

[15] Workskills curricula, Math Problem Solving, Trigonometry and Calculus are not covered.

[16] Reading comprehension (levels 3-14) has its own unique content: 3 passages, 1 randomly selected. Each passage is accompanied by 5 questions.

	Fastrack	Skills Inventory	Course Level Assessment	Module Mastery Tests
8. *Prescription*	Testlets passed set exemptions of corresponding modules	Testlets passed set exemptions of corresponding modules	Testlets passed set exemptions of corresponding modules	Passing can exempt from module if instructor allows
9. *How is start & end point set?*	Level E (default), or set by instructor[17]	Teacher assigns level A-D	All students take whole test	All students take whole test
10. *Reporting*	"Grade Level"[18]: initial, current, gain; time on task; PLATO module exemptions	Completion Module exemptions	Completion Module exemptions	Total % right; Started, completed, mastered, Time on task, # tries
11. *Look & feel*	Varies by version	Themes from courseware provide motivation & reward. Completing subtests completes a picture.	WinPLATO©, WebPLATO©	Corresponds to course
12. *Platform*	LAN/Pathways, Browser/PWLN	LAN/Pathways, Browser/PWLN	LAN/Pathways, Browser/PWLN	LAN/Pathways, Browser/PWLN
13. *What alignment used?*	FASTRACK alignments	Standard PLATO curriculum	Published tests: Published alignments User-defined tests: user-defined alignment	PLATO Tutorial modules
14. *Customization Options*	None	None	User-defined tests based on user-defined paths	Instructor assigns each test, limits tries, requires mastery
15. *Purpose*	Placement into PLATO© secondary/adult courseware	Placement into PLATO elementary courseware	Cumulative tests for PLATO secondary/adult curricula: pretest (for evaluation and placement), end-of-semester/year test	Module pre- & post-test; used to control progress (exemption, restudy, proceed to next module)

- Fig. 5: PLATO Courseware Testing Systems

[17] If reading comprehension is administered first, it resets the starting level for reading skills & strategies

[18] Grade levels refer to positions in the standard PLATO curriculum sequence. They are not comparable to norm-referenced test grade levels.

The FASTRACK Placement Testing System

The FASTRACK placement testing system is intended for entry-level testing in situations where little is known from previous tests or past instructional experience about a learner's initial skill levels in reading, writing or basic mathematics. FASTRACK emphasizes minimum testing time, minimum frustration from items which are too hard or too easy, and the convenience of individualized placement directly into PLATO tutorial curricula.

To keep the test length to a minimum, FASTRACK uses an adaptive, tailored testing algorithm, so the test items learner depends on their performance on earlier questions. Each test is organized into *testlets* of 4 questions each. Each testlet corresponds to a particular PLATO© module mastery test, and draws its items randomly from the same item pool as the module mastery test.[19]. Testlets are organized into a leveled sequence which follows the published PLATO curriculum structure. The number of testlets per level depends on the structure of the corresponding curriculum alignment. The learner starts at a level defined by the instructor (or set by a default). If the learner passes the testlets on a level, the system steps up one level and administers the next level's testlets. If the learner fails the level, the system steps down two levels and administers the testlets for that level. This process continues until the system finds the learner's skill level. Then, the system exempts the learner from the modules below that level, and the learner starts study of tutorial modules at the tested level. Thus, FASTRACK does not make judgments based on performance on a single test item, as is sometimes the case in "adaptive" tests.

To further keep testing time to a minimum, FASTRACK terminates testing for each testlet as soon as pass/fail decision can be made. Thus, if a learner succeeds on the first three questions, the fourth question in the testlet is not administered and the learner receives credit for the testlet. Similarly, as soon as the learner gets two questions wrong, the testlet is terminated and the learner fails the testlet.

The same "early termination" philosophy applies for each level as a whole. FASTRACK includes one testlet for every module mastery test. The number of modules per level depends on the structure of the corresponding PLATO curriculum, but can range from 5 to 20. Fig. 6 on the next page shows the details.

[19] Except in secondary reading, where the test items are unique to the test.

	FASTRACK		FASTRACK ADVANTAGE	
Language Arts	A 5	F 20	A 5	F 20
	B 9	G 14	B 9	G 14
	C 9	H 9	C 9	H 9
	D 13	I 7	D 13	I 7
	E 13	Total 99	E 13	Total 99
Reading Skills and Strategies	NONE		B 22	
			C 10	
			D 9 (+ 1 not tested)	
			F 9 (+ 1 not tested)	
			I 19 (+ 1 not tested)	
			Total 69	
Math	A 9	F 23	A 14	G 10
	B 12	G 14	B 17	H 14
	C 10	H 12	C 11	I 13
	D 10	I 7	D 14	J 18
	E 14	Total 111	E 14	K 20
			F 15	L 18
				Total 178

- Fig. 6: Number of modules and testlets per FASTRACK level

If 80% of the testlets at a level are passed, then testing terminates and credit is given to all testlets at that level (and the system goes on to the next higher level). On the other hand, as soon as 21% of the testlets at a level are failed, then testing at that level terminates (and the system drops down two levels and continues testing).

Levels in FASTRACK refer to the PLATO© published curriculum structure. They are expressed as letters A through M (not all curricula cover all these levels).[20] The level structure is consistent with the competency-based design philosophy of the entire PLATO© system, but there is no simple relationship between FASTRACK levels and the grade levels reported by norm-referenced tests. Educators should also carefully evaluate the meaning of a reported FASTRACK level, relative to their own curriculum structure. For example,

[20] Reading Comprehension reports grade levels based on the difficulty of the passages used, rated by the Flesch-Kinkaid formula.

suppose FASTRACK places a learner at Level D in mathematics. Only if the learner's school teaches the Level D skills in fourth grade would it be appropriate to place the learner in the school's fourth grade curriculum.

For convenience, FASTRACK also reports starting, in progress, and ending "grade level" as the learners work through additional modules in the PLATO courseware. The same competency-based considerations apply in interpretation of these gains: they cannot be simply interpreted as equivalent to, or predictive of, the grade level scores on a norm-referenced test.

The combination of tailored testing and early testlet termination means that the total length of a FASTRACK test can vary considerably from one learner to the next. For example, suppose a learner is actually at level K, but the teacher starts FASTRACK testing at level D. Then FASTRACK will administer the testlets for levels D through L before placing the learner at level K. For example, suppose further that there are 20 testlets per level. The learner will see at least three questions per testlet, and the total test length will be between 271 and 360 questions, depending on how well the learners does on each testlet. Suppose, however, that the learner is actually at level B, and the instructor starts testing at level D. In that case, testlets and level testing at levels D will terminate early and the test will skip to level B. The learner might see as few as 60 items.

At times, the adaptive/prescriptive testing in FASTRACK can lead to substantially different conclusions about a learner's level than would emerge from a norm-referenced test. There are three reasons for this: first, as explained above, FASTRACK is a competency-based system and its levels do not necessarily correspond to the grade levels in a norm-referenced system. Second, FASTRACK includes testlets corresponding to every module mastery test in the corresponding PLATO curriculum. Thus, it does not engage in domain sampling, but rather attempts to take a full inventory of skills at a given level – if the learner is succeeding at that level. On the other hand, the early termination provision means that as soon as the learner shows a significant pattern of failure at a given level, testing at that level is terminated and the test drops down two levels. Third, FASTRACK is often used in remedial situations. It is characteristic of these situations that the learners have very irregular skill profiles: they may master some subskills at a variety of levels, but not others. In this case, FASTRACK will usually refuse to give them credit for the entire level (unless the deficits are in fewer than 20% of the testlets at a level). These three reasons combine to mean that FASTRACK can produce a fairly conservative estimate of a learner's standing, in comparison to the normed levels from a standardized test. Since FASTRACK's purpose is placement, however, this is appropriate: learners should start with what they know, and if the instructor allows, they can use the courseware module mastery tests to "place out of" modules they know.

However, the design of FASTRACK as a placement system means that it is not recommended for pretest/posttest use in program effectiveness reports and evaluations. When used as a posttest, the possibility of a conservative estimate of skills works against a true assessment of the learner's progress. Furthermore, research has shown that a reliable measure of mastery of a given objective requires eight homogeneous test items on that objective. This is well above the testlet length in FASTRACK. Instead, we recommend use of the Course Level Assessments, or tests provided by *Orion* GATE or *PLATO*© *LINK*, for these purposes.

FASTRACK provides individualized prescriptions into PLATO courseware. It does this by sending an "exempt" code to the management system for each module for which the learner has passed the corresponding FASTRACK testlet, and for all modules in a level in which the learner has passed 80% of the testlets. Functionally, the management system treats the "exempt" code as if the learner had passed the full module mastery test. However, for clarity, some detail reports generated by the management system distinguish between exemption and mastery, depending on the purpose of the report.

In sum, the FASTRACK system provides a convenient, efficient way to perform a competency-based assessment of incoming learners, and to place them in PLATO courseware. FASTRACK systems are available for most core tutorial curricula in PLATO courseware. For details of availability, refer to the current product information.

Elementary Skills Inventory

For placement testing in certain elementary curricula, PLATO Learning offers the Elementary Skills Inventory. Tests are published for the entire elementary mathematics curriculum, and for *Elementary Reading for the Real World.* Instructors may assign the tests for individual courses in any curriculum.

Items in the Reading and Math Inventories closely parallel the learning content, format and style contained in the corresponding PLATO courses. Test items are tailored to match the specific learning activities in the PLATO courses, and to mirror the themes, characters and general vocabulary used in the courses. In this way, topics are tested as PLATO Learning teaches them, and continuity is maintained for the students. All items are in multiple-choice format with one correct response. Interest and variety have been added to the items by use of color, sound, graphics and a variety of item layouts.

Note, however, that the Inventory item pools are completely separate from those used in the courseware. A total of 2,300 items are available: 900 items for reading, and 1,400 items for math. The items have been pilot-tested with actual learners and analyzed using standard item analysis statistics.

The tests are organized into a sequence of testlets, one per courseware module. Each testlet randomly draws 5 items from a pool of 10. Note, however, that the items are unique to the test and are not used in module mastery tests. Furthermore, this system does not use the tailored testing algorithm of FASTRACK. However, it does terminate testing on any testlet if the learner gets 4 items right or 2 items wrong.

A common problem with very young children is to maintain their interest and motivation throughout a test. The Elementary Skills Inventory has a unique solution. Based on the appeal of stickers for young children, the testing system awards "electronic stickers" for completion of each subtest. When all stickers are received, an entertaining graphic is complete. Furthermore, the general interface has a game-like quality designed especially to be appealing to these young learners. The tests can be interrupted and resumed later.

The system also includes a training module to teach students the interface before beginning the test. This improves test reliability.

The system determines a grade level for each student, and places the student in the corresponding PLATO courseware while exempting them from modules which are not needed. Thus, the tests set exemptions in the corresponding modules from the PLATO elementary curricula, and a summary report of completion and mastery is available.

To shorten testing time, the instructor can set a starting grade level. The test will then adjust duration based on the student's skill level. The tests also use an early termination algorithm for each testlet. If students miss two items for an activity, no more problems will be given and the program will proceed to the next testlet. If students answer four items correctly, the system will not offer the fifth, since mastery has been demonstrated. Thus, mastery of each testlet is set at 80%.

Module Mastery Tests

All PLATO tutorial courseware uses a modular structure. Modules generally are sized to be completed in approximately 30-45 minutes (with considerable variability due to learner abilities). Most modules teach a single terminal objective, with 5-9 enabling objectives. Most modules consist of three lessons: tutorial, application/practice, and mastery test.

Module mastery tests can be used before or after study of their corresponding tutorial and practice lessons. If you allow it, learners can take the test before study, and if they pass, they will be allowed to skip the tutorial. After the learners study each tutorial, you can require learners to pass the mastery test before going on to the next module.

The module mastery tests are at the heart of all PLATO courseware testing facilities. The module mastery tests typically consist of 5-10 items drawn randomly from a pool of items which is three times larger than the test length: a total of over 30,000 items in all modules. These item pools are also used to generate the FASTRACK placement tests and the Course Level Assessments. To limit item exposure (and attempts to "game the system" by repeatedly taking the mastery tests until all of the items have been memorized), we recommend that you limit the number of allowable tries on all tests to two or three. You can set this option in the management system.

Most modules have a single terminal objective. Most tests are designed to be a homogeneous measurement of the module's terminal objective. Tests of enabling objectives are included in exceptional cases if the instructional designer determines that it is not reasonable to assume that proficiency on the terminal objective includes proficiency on all the enabling objectives. If the module has more than one terminal objective, then there are usually separate module mastery tests for each objective. Thus, some modules have more than one mastery test.

Content Validity

Content of the module, including the mastery test, is based at the high level on a careful analysis of all state and national standards and benchmarks, using the synthesis incorporated in the PLATO© standards data base and the synthesis maintained by the Mid-Continent Regional Educational Laboratory (McREL). At a more detailed level, content is based on a rigorous task/content analysis of the knowledge structure underlying the standards.

This two-level analysis assures there is a high degree of content validity as module mastery tests refer to standards. Determining which module mastery test corresponds to your state's standards and benchmarks is simply a matter of referring to the current PLATO courseware standards alignments for your state or local standards. These alignments are published by PLATO Learning, or you can generate them by querying the *Orion* or the *Polaris* data base.

Item Formats

As explained in Part 2, the PLATO courseware item banks are constructed based on test specification matrices which include model items of each type required and number of items to be generated. Most items are written by hand, but certain kinds of mathematics items are generated by algorithm.

PLATO Learning's tutorials and application/practice lessons use a mix of question formats including extensive use of constructed response (short answer or completion) formats, specialized computer formats (such as drag and drop construction or graphic identification interactions), and multiple choice. In module mastery tests, however, reliability considerations make it particularly important to use item formats which are consistent with each other and pose no obstacle to accessibility to those with limited keyboard skills. Consequently, item formats are limited to multiple choice and very simple-to-use short answer formats, such as those requiring entry of numerals.

Most mastery tests randomly select individual items from their item pools. However, in certain specialized cases, testlets of items in a fixed sequence are randomly selected, instead of individual items. This practice is used in two circumstances: first, if the test is of procedural knowledge, where it is desirable to test the learner's ability to perform individual steps in the procedure, then the designer will construct a series of items corresponding to the key steps and decisions in the procedure. Second, in tests of reading comprehension, testing for comprehension of a long reading passage requires more than one item for each passage.

Scoring and Reliability

PLATO courseware item pools are constructed using the rigorous design and review process described in Part 2. Standards applied reflect current standard professional practice, as described there.

Module mastery tests typically select 5-10 items from their pools. Since the items generally are a homogeneous measure of the module's terminal objective, this number of items is in the recommended range of test length for reliable measurement of one objective (most sources recommend 7-8 items per objective). In current and future new development and upgrades, all item banks are expanded to support test lengths of 10 items.

The criterion for mastery of each test is fixed at 80% (4 out of 5, 8 out of 10, etc.). As explained in Part 2, no test is a perfect measure of its content, so it is not reasonable to set mastery at 100%.

To keep tests short and minimize frustration, the tests terminate early if the learner reaches the mastery score, or if the number of items wrong will preclude mastery (more than 20% wrong).

Reports

The PLATO© management system stores test score (as a percent), mastery status (not started, started, complete, mastered, or exempt), number of tries, and total time on task for each module mastery test. These data are reported in a wide variety of formats, including easy-to-read bar graphs and detailed numerical reports, for individual students and groups.

It is very useful to examine all four types of data, when judging the performance of individual learners. For example, a learner who is floundering or who is trying to "game the system" will usually show total time on task for the tutorials to be much less than typical for most learners in the class, and may show an abnormally high number of tries. You can use the exception reporting features of the management system to identify these learners quickly.

Since the module mastery tests are competency-based, they are best used to make a binary "mastered/non-mastered" decision on each competency. If it is necessary to use module mastery tests to assign grades, we do not recommend use of the scores on each test as a grade if learners are allowed to take the test repeatedly, as is the case in mastery learning contexts. A preferred method of grading can be based on number of modules mastered in a given time period. Alternatively, percent of modules mastered versus number assigned can be used to derive a grade. Any grading method depends on policies which must be established by the instructor and the school.

Course Level Assessments

PLATO curricula often are divided into courses. In larger curricula, the courses may contain a few dozen modules each. Pre-defined course-level assessments (CLA's) are cumulative tests for each course, using the original curriculum sequence of the course or a published state standards alignment. CLA's are available for the secondary writing and mathematics curricula, and are included in many published alignments.

For clients who have constructed their own learning paths within the PLATO system, it is possible to generate CLA's based on this custom learning path. Custom CLA's allow the instructor to specify which modules to test, number of items to use, and cutoff score for passing. This capability is particularly useful for clients who need standards-aligned cumulative tests.

CLA's are intended for pretest/posttest use, to measure gain, retention, and overall level of achievement in a defined block of instruction. When used as a pretest, the CLA will set exemptions to the corresponding modules (when used as a post-test, exemptions are still set, but for modules which have already been studied, so the exemption information has no meaning). Since CLA's draw randomly from the module test item pools, item exposure from pretest to posttest is low.

CLA's are often a preferred method of deriving a course or unit grade from PLATO study. Since the instructor can define cutoff scores, they can be set to correspond to grading standards.

Published CLA's include 4 items per corresponding module mastery test. Note that, unlike FASTRACK and module mastery tests, CLA's do not use early termination algorithms, so all learners see all items on the test. Note also that, unlike FASTRACK, CLA's do not use a tailored testing algorithm. Thus, using CLA's for placement avoids the issues of interpreting placement levels which are implicit in the FASTRACK design. On the other hand, CLA's may be longer for some learners than the analogous FASTRACK test, particularly if the scope of curriculum to be tested is large.

Part
5

Choosing the Right PLATO© Test for Your Needs

With all the testing options PLATO provides, clients may find daunting the task of choosing the right test for their needs. Use the questions below to help sort out the options.

1. **Is my test high stakes or low stakes?**

    - If the test low stakes, use PLATO tests

    - If the test is high stakes, use a competency-based state standards tests. If none is available, or if your state requires, use a norm-referenced (standardized) test.

2. **Is my test used to support the general curriculum, or is it to practice for a high-stakes test such as a state standards test, or is it in support of study of PLATO courseware?**

    - If the purpose is general, use *Orion* GATE or *PLATO LINK*

    - If the purpose is practice for a high-stakes test, use the *Simulated Test System.*

    - If the purpose is study of PLATO courseware, use the PLATO courseware tests.

3. **What platform am I using?**

    - If your system is browser-based with Internet access (Windows, MacIntosh), use PLATO© *Orion*, *PLATO© LINK*, or PLATO© Web Learning Network.

    - If your system is a local area network (LAN) or Intranet, use the PLATO© Pathways management system and PLATO courseware tests.

    - If you wish to use paper-and-pencil tests, and you have Internet access, use *Orion* or *PLATO LINK©*.

- If you wish to use paper-and-pencil tests, and you do not have Internet access, check to see if *Polaris* is available for installation on your internal network.

4. **What Curricula am I testing?**

   - Math: all tests
   - Reading: all tests except CLA
   - Writing: *Orion* GATE, *PLATO© LINK*, PLATO courseware tests, National Writing Test

5. **What Level am I testing?**

   - K-3: Skills Inventory, Module Mastery Tests
   - 3-12: *Orion* GATE, *PLATO© LINK*
   - Secondary: Plato secondary courseware tests, *Simulated Test System*

6. **For What Purpose am I testing?**

   - Placement into PLATO courseware– Elementary : Skills Inventory
   - Placement into PLATO courseware – Secondary: FASTRACK or CLA
   - Cumulative Pre/Post Tests: Orion GATE, *PLATO© LINK, CLA*
   - Practice for Standards Test: *PLATO© LINK Simulated Test System*

7. **I want to use my own test items, as well as those provided, in a variety of item formats.**

   - PLATO© *Orion* GATE

8. **I do frequent retesting, and I want to minimize item exposure.**

   - PLATO LINK©

9. **I want to create tests which support my textbook chapters, as well as standards.**

   - PLATO LINK©

10. I want immediate, diagnostic reports and prescriptions to:

- PLATO Courseware: all testing systems
- My district's textbooks, screened web resources : *Orion* GATE
- Web resources, textbooks: PLATO© LINK

11. I want real-time reporting to:

- Students, Teachers, Administrators: all systems
- Parents: *Orion* GATE; PLATO© LINK

12. I want:

- Just the testing: PLATO© LINK, Simulated Testing System
- Full integration of resource planning, curriculum planning, testing: *Orion* GATE
- Just testing and courseware: PLATO courseware tests

Finally, when implementing a testing system, it is important to plan for professional development which addresses not only the "buttonology" of the new software, but also the effective construction, interpretation and use of the new system for purposes of placement, progress control, and cumulative post-testing. Only when testing is a fully integrated tool does it become useful for data-based decision making, rather than being an extra "layer" of irrelevant tasks, for teachers implementing standards.

References

American Educational Research Association., American Psychological Association., et al. (1999). Standards for educational and psychological testing. Washington, DC, American Educational Research Association.

Bloom, B. S. (1982). Human characteristics and school learning. New York, McGraw-Hill.

Gronlund, N. E. (1993). How to make achievement tests and assessments. Boston, Allyn and Bacon.

Haladyna, T. M. (1997). Writing test items to evaluate higher order thinking. Boston, Allyn and Bacon.

Kendall, J. S. (2001). A Technical Guide for Revising or Developing Standards and Benchmarks. Colorado, Mid-continent Research for Education and Learning, 2550 S. Parker Road, Suite 500, Aurora, CO 80014-1678. Tel: 303-337-0990; Fax: 303-337-3005; Web site: http://www.mcrel.org.

Linn, R. L., National Council on Measurement in Education., et al. (1993). Educational measurement. Phoenix, AZ, Oryx Press.

Marzano, R. J. and J. S. Kendall (1996). The Fall and Rise of Standards-Based Education. Issues in Brief. Colorado, NASBE Publications
National Association of State Boards of Education, Alexandria, VA.

Marzano, R. J. and Mid-Continent Regional Educational Lab. Aurora CO. (1998). Models of Standards Implementation: Implications for the Classroom. Colorado: 87.

Osterlind, S. J. (1998). Constructing test items : multiple-choice, constructed-response, performance, and other formats. Boston, Kluwer Academic Publishers.