Research Paper

Benchmark Assessment in Standards-Based Education:

The Galileo K-12 Online Educational Management System

by
John Richard Bergan, Ph.D.
John Robert Bergan, Ph.D.
and Christine Guerrera Burnham, Ph.D.

Assessment Technology Incorporated

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

Assessment Technology, Inc., Publishers
Tucson, Arizona, U.S.A.
Printed in the United States of America.
V6-061711

Benchmark Assessment in Standards-Based Education

By John Richard Bergan, Ph.D., John Robert Bergan, Ph.D.,
and Christine Guerrera Burnham, Ph.D.
Assessment Technology, Incorporated

Table of Contents

Acknowledgements

The authors wish to thank David Thissen, Jason Feld, Kathryn Bergan, and Jody Jepson for their thorough review of this paper and for their helpful comments, which led to improvements of the manuscript.

We also wish to thank the students, teachers, and administrators in the participating school districts for their efforts, which made it possible to present the data on benchmark assessment included in the paper.

John R. Bergan, Ph.D.
President, Assessment Technology Inc.

Introduction: Benchmark Assessment in Standards-Based Education

It is well known that the technological achievements of the 20th century have produced global societal changes that are posing new challenges for education around the earth. The world's knowledge base is expanding at an unprecedented rate. The number of people who can easily access information has expanded exponentially and the speed at which information can be obtained has accelerated to a degree that could hardly have been imagined a few short decades ago. The world is growing smaller and more competitive. As a result, we know that the requirements for an educated citizenry in the years ahead will rise to unprecedented levels. Those who can learn rapidly, can absorb and organize information efficiently, and can use what is known in creative ways will prosper and advance in the global community. Those who lack the skills to take advantage of the opportunities afforded by the information age will face the danger of being left far behind.

In response to the challenge of preparing the young to succeed in the new age, educators across the nation are exploring new ways to increase the overall capabilities of today's students. Advances in the management of education lie at the heart of these explorations. Benchmark assessment has become a significant tool in the management process. The role of benchmark assessment is to provide to students, teachers, parents, and administrators the timely information needed to plan opportunities to promote learning.

Assessment Technology, Incorporated (ATI) supports benchmark assessment and its implementation in standards-based education through the Galileo Educational Management System (EMS) implemented in Galileo K-12 Online. Galileo K-12 Online includes a variety of assessment and curriculum tools designed to inform instruction. The assessment tools in the system include capabilities to construct, administer, and score benchmark assessments, formative assessments, diagnostic assessments, and district-wide examinations such as end-of-course examinations. The assessment tools allow teachers and administrators to construct tests from items selected by searching ATI item banks. There are also tools that enable teachers and administrators to build their own test items, and there are tools that make it possible to import district-made tests, print them, and score them online. Tests can be administered online or offline. Student responses to offline assessments can be scanned using a plain-paper scanner or high-speed scanner. Multiple-choice and true/false items can be scored automatically and reports can be generated immediately following scoring. Curriculum tools, detailed in other documents, make it possible for teachers to create unit plans, lesson plans, and assignments online and to publish lesson plans and assignments online for students.

The benchmark assessment tools in Galileo deserve special consideration because benchmark assessment is a relatively new form of assessment and because this form of assessment can play an important role in standards-based educational initiatives. The standards-based approach to education, which has swept the nation in recent years, offers new opportunities and carries new responsibilities for local educational agencies as they work to attain unique local goals as well as shared goals articulated in state standards. Benchmark assessment plays a key role in providing information to assist educational agencies to achieve their goals.

This paper describes benchmark assessment and its role in standards-based education. It details the kinds of information provided through benchmark assessment and how that information can be interpreted. Finally, it describes key reports available in Galileo K-12 Online

and discusses the ways in which benchmark results conveyed through those reports can be used in making decisions to guide instruction.

I. Benchmark Assessments: Characteristics, Uses and Psychometrics

The discussion in this section describes the defining features of benchmark assessment. These include the link between benchmark assessments and standards and the relationship between benchmark assessment and instruction. The discussion also focuses on the uses of benchmark assessment in determining the mastery of standards and in identifying students who may be at risk for not mastering standards. The discussion closes with a brief outline of the analyses required to establish the psychometric properties of benchmark assessments.

A. Standards-Based Assessment

A benchmark test is a locally customized, district-wide assessment designed to measure the achievement of standards. The first step in constructing a benchmark test is to identify the learning standards or performance objectives that the assessment will measure. Standards and performance objectives are generally those specified at the state level. However, local standards may also be included. Learning standards or performance objectives selected to guide the construction of a benchmark test are entered online in the *Assessment Planner.* The *Assessment Planner* automatically generates a benchmark test customized to district specifications. The automatically generated test may be reviewed online. Following review, the final version of the test is published online and made available for online or offline administration to students.

B. Assessment to Inform Instruction

The fundamental purpose of benchmark assessment is to provide information that can be used to guide instruction. Benchmark tests measure student mastery of standards targeted for instruction. In so doing, they indicate what students have accomplished when given appropriate learning opportunities. Benchmark tests also inform instruction in cases in which standards have not been mastered even though appropriate learning opportunities have been provided. Schools are now initiating interventions that go beyond the basic responsibility of providing appropriate learning opportunities. These interventions focus not just on whether skills have been taught, but rather on whether or not they have been learned. Benchmark tests support these types of interventions by pinpointing the specific skills that students need to acquire to master standards that have been targeted for instruction, but that have not yet been met.

Benchmark assessment provides information to guide instruction in cyclical fashion. In some cases initial instruction is preceded by a pretest designed to provide an overall picture of initial student mastery of standards. Initial instruction is followed by a benchmark test designed to assess mastery of standards covered during the initial instructional period. For example, a benchmark test may be administered after instruction implementing a unit plan has been completed. Teachers and administrators may use the results of the assessment to plan and implement interventions to address areas in which students may not have displayed mastery of standards and performance objectives measured on the test. For instance, a reteaching intervention may be employed assisting students to master standards that they have not yet

met. Short formative tests may then be used following reteaching to ensure that standards not initially mastered have been mastered following reteaching.

C. Multiple Assessments of Standards Mastery

Typically the cycle of teaching, assessment, and intervention will be implemented three or four times during the school year. Repetitions of the cycle provide an increasing body of information about student learning. This information coupled with information on statewide test performance provides the opportunity for a multi-test approach to the assessment of standards mastery.

It is often useful to base decisions about standards mastery on multiple tests rather than a single assessment. The multi-test approach is particularly useful when high-stakes decisions are linked to judgments about the mastery of standards. Achievement test results invariably include some amount of measurement error. When high-stakes decisions are made on the basis of a single test, there is an implicit assumption that the test in question measures student proficiency without error (Bergan, Bergan, & Guerrera, 2005). By contrast when multiple tests are used to determine the mastery of standards, each of the tests may be assumed to be a fallible indicator of student performance. Moreover, the extent of classification errors based on multiple assessments can be made explicit.

In addition to the well-recognized fact that the multi-test approach can reduce the impact of a single test on high-stakes decisions, the approach also increases the likelihood that the assessments used to assess mastery cover the full range of content that has actually been taught. This issue has been a long-standing concern in the design of accountability initiatives (National Research Council, 1999). Finally, the multi-test approach increases timely access to assessment information that counts in the overall determination of mastery.

The use that can be made of the multi-test approach depends on state and federal regulations. One important way that districts using Galileo K-12 Online have used multi-test information is in providing evidence of standards mastery in cases in which other available evidence is insufficient or subject to question.

D. Risk Assessment

When a relationship has been established between performance on one or more benchmark assessments and performance on a statewide test, benchmark results can be used to assess the level of risk that a given student will not meet state standards as measured by the statewide test. The probability of accurately forecasting mastery of state standards will depend in part on the strength of the relationship between each benchmark assessment and the statewide test and in part on the mastery patterns evidenced by the students. For example, consider the situation in which performance on each of three benchmark tests is correlated with performance on a statewide test. Suppose that a student has failed to meet standards on all three benchmark tests. The probability that the student will meet standards as measured by the statewide test will in all likelihood be substantially lower than will be the case for a student who has met standards on all three benchmark assessments. Table 1 illustrates the type of information that may be used to inform risk assessment. After data of the type shown in the table have been gathered, the results can be used to forecast risk given subsequent administrations of benchmark tests. As additional data become available, risk forecasts can be adjusted.

TABLE 1
Risks of not Meeting State Standards Given Varying Patterns of Benchmark Standards Mastery

Assessments	District Risk Assessments - 5th Grade Math Benchmark Mastery Patterns 1	2	3	Number of Students	AIMS Test Met	Not Met
Risk Assessment 1						
Benchmark 1	Met			417	0.94	0.06
	Not Met			44	0.41	0.59
Risk Assessment 2	Met	Met		395	0.96	0.04
Benchmarks 1,2	Met	Not Met		22	0.59	0.41
	Not Met	Met		21	0.57	0.43
	Not Met	Not Met		23	0.26	0.74
Risk Assessment 3	Met	Met	Met	375	0.98	0.02
Benchmark 1,2,3	Met	Met	Not Met	20	0.50	0.50
	Met	Not Met	Met	14	0.79	0.21
	Met	Not Met	Not Met	8	0.25	0.75
	Not Met	Met	Met	13	0.77	0.23
	Not Met	Met	Not Met	8	0.25	0.75
	Not Met	Not Met	Met	5	0.80	0.20
	Not Met	Not Met	Not Met	18	0.11	0.89

Table 1 summarizes data from three benchmark assessments and a statewide test, the Arizona's Instrument to Measure Standards (AIMS). The first benchmark test was administered in the fall, the second in the winter, and the third in the spring. The AIMS test was administered toward the end of the school year following the three benchmark tests. The results of three risk assessments are shown in the table. The three assessments were designed to provide information about risk given data available at different points in the school year. For example, the first opportunity to assess risk using benchmark assessment occurs after the administration of an initial benchmark test. At this point the only benchmark information available for risk assessment comes from one test. By contrast, toward the end of the school year, information from multiple benchmark tests may be used to assess risk. For instance, in the example given here, risk assessment could utilize information from three benchmark tests.

The first risk assessment shown in Table 1 examined the risk of not meeting the state standard as measured by AIMS given that the benchmark standard was met or not met on Benchmark 1. The second assessment examined the risk of not meeting the state standard given the various possible mastery classifications for Benchmark 1 and Benchmark 2. The third risk assessment utilized information from all three benchmark tests.

Results for the first assessment show that 417 students were classified as meeting the state standard and 44 students were classified as not meeting the standard based on their performance on Benchmark 1. The probability of meeting the standard on the statewide test given that the standard was met on Benchmark 1 was .94. The probability of not meeting the standard on AIMS was only .06. On the other hand the probability of not meeting the standard on AIMS for students who did not meet the standard on Benchmark 1 was .59. Thus, the risk of not meeting the standard on AIMS was much higher for students who did not meet the benchmark standard than for those who did meet the benchmark standard.

The results for risk assessment 2 and risk assessment 3 show that the risk of not meeting the state standard increases when students fail to meet the standard on consecutive benchmark tests. The probability of not meeting the standard on AIMS was .74 for students who failed to meet the standard on both Benchmark 1 and Benchmark 2. For students who failed to meet the standard on all three benchmark tests, the probability of not meeting the standard on AIMS was .89. These students had approximately one chance in ten of meeting the state standard.

Risk assessment information of the type provided in this example has implications for the design of interventions aimed at assisting students to meet standards. The results suggest an approach to intervention planning that takes account of information about risk that increases with each benchmark assessment. As results from each benchmark become available, intervention plans can be adjusted based on all of the available information. For example, after Benchmark 1, an extended intervention could be designed for students identified as at risk for not meeting the state standard. Toward the end of the school year, groups of students may be identified who are at high risk for not meeting state standards. Intensive short term interventions may be designed to meet the needs of those students.

E. Benchmark Psychometrics

Standards-based educational systems being implemented in states across the nation are designed to enable local educational agencies to pursue common goals reflected in state standards. Statewide tests provide a measure of the achievement of the goals reflected in standards. Benchmark tests used to guide instruction also provide measures of the achievement of the goals reflected in standards. It is essential that there be commonality in what is measured on benchmark tests and the statewide test to insure that benchmark measures used to guide instruction are assessing the same capabilities as those reflected in the statewide test. For example, if benchmark tests are to be used for risk assessment and/or multi-test determination of standards mastery, there must be commonality between what is assessed on local benchmark tests and the statewide test. The necessary commonality between benchmark tests and statewide tests requires that psychometric analyses be conducted on benchmark tests just as they are on statewide tests.

i. Item Parameter Estimation

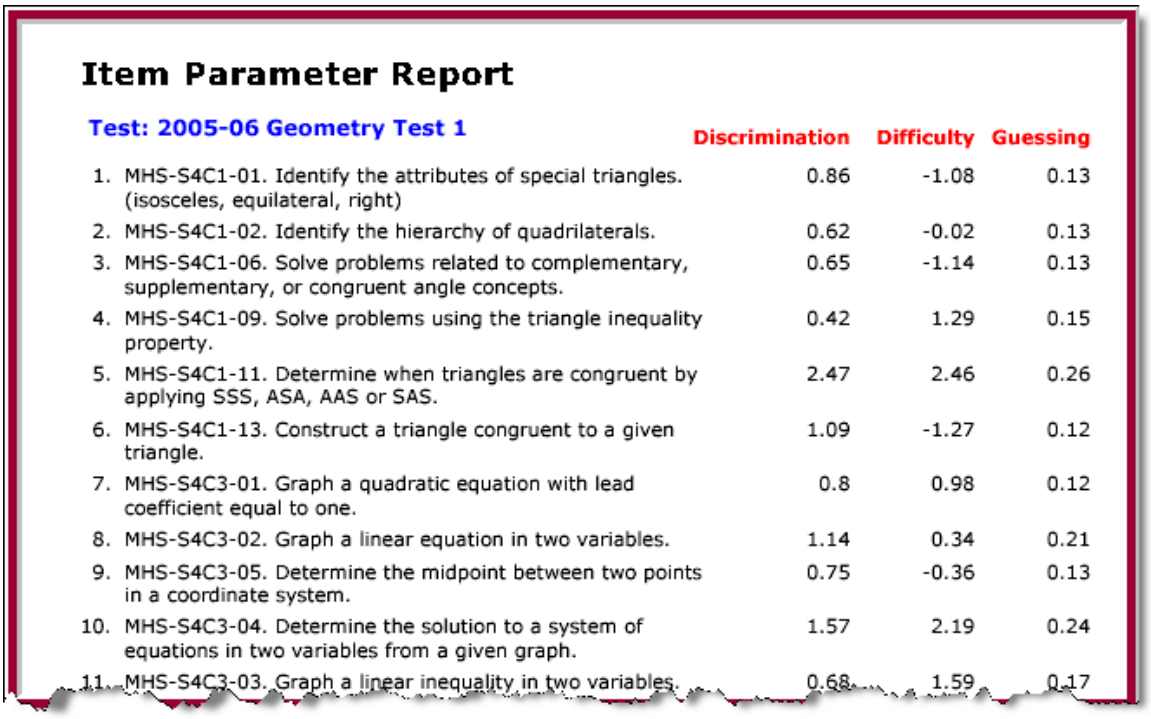
The first step in establishing benchmark psychometrics is to examine the psychometric properties of items included in the test. ATI conducts this type of examination for benchmark assessments using Item Response Theory (IRT). IRT assumes that a student's response to a test item is determined by the student's ability and certain item parameters, i.e., characteristics of the item. For multiple-choice tests, ATI uses an IRT model including three item parameters: a discrimination parameter, a difficulty parameter, and a guessing parameter. The IRT model is used to estimate the values of these three parameters for each item on the test. The discussion that follows describes each of the three parameters:

1. Discrimination Parameter - indicates the extent to which an item discriminates sharply between different levels of ability. Values approaching or exceeding 1.0 discriminate between levels of ability very well. Values close to 0.0 discriminate between different ability levels very poorly. The discrimination parameter indicates the relationship of the item to the underlying ability being measured. Items with high discrimination values make a positive contribution to the test reliability.

2. Difficulty Parameter - provides information on the relative difficulty of items in a test. In general, it is useful to construct tests including a broad range of difficulty levels. Tests of this kind will be sensitive to a range of ability levels. Such tests generally correlate higher with criterion measures (e.g., statewide assessments) than do tests sensitive to a limited range of ability levels. Zero is the average ability of the students. An item difficulty of zero is of appropriate difficulty for the average student. If item difficulty is above zero, the item is more difficult. When item difficulty is negative, the item is less difficult.

3. Guessing Parameter - indicates the likelihood a student who does not know the answer to the question posed in a multiple-choice item will guess the correct answer. Given a multiple-choice item with four alternative choices, it would be reasonable to expect that the chances of guessing the correct answer would be about one in four, or .25. Sometimes this will be the case. However, sometimes the probability of guessing the correct answer will turn out to be lower than .25 and sometimes it will be higher than .25. Items that make it easy to guess the correct answer are less desirable than items that limit the likelihood of guessing the answer correctly.

Information regarding item parameter estimates for benchmark tests is provided through an item parameter report available in Galileo K-12 Online. Figure 1 shows a sample item parameter report.



Item Parameter Report

Test: 2005-06 Geometry Test 1

	Discrimination	Difficulty	Guessing
1. MHS-S4C1-01. Identify the attributes of special triangles. (isosceles, equilateral, right)	0.86	-1.08	0.13
2. MHS-S4C1-02. Identify the hierarchy of quadrilaterals.	0.62	-0.02	0.13
3. MHS-S4C1-06. Solve problems related to complementary, supplementary, or congruent angle concepts.	0.65	-1.14	0.13
4. MHS-S4C1-09. Solve problems using the triangle inequality property.	0.42	1.29	0.15
5. MHS-S4C1-11. Determine when triangles are congruent by applying SSS, ASA, AAS or SAS.	2.47	2.46	0.26
6. MHS-S4C1-13. Construct a triangle congruent to a given triangle.	1.09	-1.27	0.12
7. MHS-S4C3-01. Graph a quadratic equation with lead coefficient equal to one.	0.8	0.98	0.12
8. MHS-S4C3-02. Graph a linear equation in two variables.	1.14	0.34	0.21
9. MHS-S4C3-05. Determine the midpoint between two points in a coordinate system.	0.75	-0.36	0.13
10. MHS-S4C3-04. Determine the solution to a system of equations in two variables from a given graph.	1.57	2.19	0.24
11. MHS-S4C3-03. Graph a linear inequality in two variables.	0.68	1.59	0.17

FIGURE 1
Item Parameter Report

ii. Reliability

If benchmark tests are to serve the purposes for which they are intended, they must be reliable. Reliability has to do with the consistency of information provided by an assessment. A particularly important form of reliability for benchmark assessment as well as other types of assessment is internal consistency. Measures of internal consistency provide information regarding the extent to which all of the items on a test are related to the underlying ability that the test is designed to measure. Benchmark tests are designed to correlate with other measures of student proficiency including statewide assessments. A test that lacks internal consistency does not correlate well even with itself. Therefore, it is unlikely that it would correlate well with other measures.

Table 2 illustrates internal consistency information for benchmark assessments. The table presents marginal reliabilities for the benchmark tests used in a study of standards mastery (Bergan, Bergan, & Guerrera, 2005). Marginal reliability coefficients are measures of internal consistency that may be easily computed in the course of psychometric analyses involving IRT. The marginal reliability coefficient combines measurement error estimated at different points on the ability continuum into an overall reliability coefficient, which corresponds quite closely to other widely used coefficients such as coefficient alpha.

TABLE 2
Marginal Reliabilities for Benchmark Tests

Test	Reliability	N
Math 1: 3rd Grade	0.95	2348
Math 2: 3rd Grade	0.92	2490
Reading and Literature 1: 3rd Grade	0.94	2326
Reading and Literature 2: 3rd Grade	0.92	2422
Math 1: 5th Grade	0.94	2587
Math 2: 5th Grade	0.94	2697
Reading and Literature 1:5th Grade	0.94	2626
Reading and Literature 2: 5th Grade	0.90	2633
Math 1: 8th Grade	0.92	2458
Math 2: 8th Grade	0.86	1850
Reading and Literature 1: 8th Grade	0.89	2794
Reading and Literature 2: 8th Grade	0.93	2107

In this example, two of the marginal reliability coefficients were in the 80s and the rest were above .90. ATI benchmark tests with reliabilities in the 80s and 90s have been used effectively in forecasting and multi-test standards mastery initiatives.

Reliability is directly affected by test length. Longer tests tend to be more reliable than shorter tests. Figure 2 plots the relationship between benchmark test length and reliability for 156 benchmark assessments administered in multiple grades and multiple school districts. The data in the figure suggest that adequate levels of reliability can be achieved with benchmark assessments that are about 35 items to 40 items long. ATI recommends that benchmark assessments contain a minimum of 40 items to ensure adequate reliability.

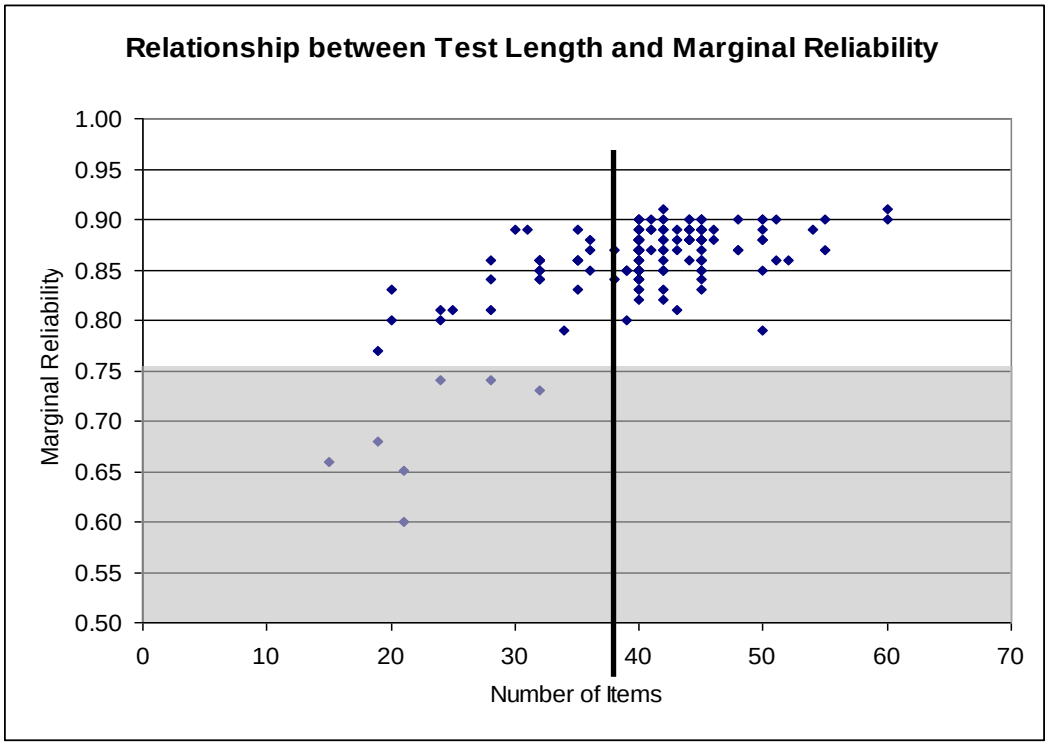


FIGURE 2
Assessments consistently begin to reach an acceptable level of reliability at a length of about 35 to 40 items.

iii. Validity

Standards-based educational initiatives across the nation are targeting instruction toward the achievement of state and local standards. State standards provide a common set of goals for all school districts within the state. Statewide tests play a critical role in measuring the achievement of state standards. Benchmark assessments are also intended to measure the achievement of state standards. Accordingly, it is reasonable to expect significant correlations between benchmark tests in a particular state and the statewide test for that state. A finding revealing such correlations would provide important evidence of the validity of the benchmark assessments.

Although significant correlations support the validity of benchmark tests, it is important to recognize that benchmark tests differ from statewide tests in significant ways. The two forms of assessment serve different purposes. Statewide tests are typically administered toward the end of the school year to provide accountability information for the state, local education agencies, and the public. Benchmark tests are administered periodically during the school year to guide instruction. The skills assessed on a benchmark test are typically selected to match skills targeted for instruction in the curriculum at a particular time. This is not the case for statewide tests. For these and other reasons, benchmark tests should not be thought of as replicas of statewide tests. Accordingly, although the two forms of assessment should correlate, the correlations should not be expected to be as high as the correlation between parallel forms of the same test.

Multiple benchmark tests administered during the school year measure student achievement in the same or related knowledge areas. As a result, it is reasonable to expect

benchmark tests to correlate well with each other. Thus, correlations among benchmark assessments provide evidence of the validity of benchmark assessments.

Table 3 illustrates the kinds of data that can be used to support the validity of benchmark assessments. The table shows a series of correlation matrices for district benchmark tests and AIMS, the statewide test used in Arizona. Each matrix reflects the relationships among benchmark tests and AIMS for a particular grade level. The selected grade levels are those at which the AIMS test was administered during the 2004-2005 school year in the elementary and middle schools in the state. The correlations between benchmark tests and AIMS ranged from .58 to .78 with a mean of .69. The relationships among the benchmark tests were slightly lower than the relationship between the benchmark tests and AIMS. The correlations among the benchmark tests ranged from .45 to .77 with a mean of .66.

TABLE 3
Correlation matrices for district math and reading tests

Correlation Matrix for Third Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.74	1.00		
3 Math Benchmark 3	.70	.73	1.00	
4 Math AIMS	.72	.73	.76	1.00

Correlation Matrix for Third Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.75	1.00		
3 Reading Benchmark 3	.71	.72	1.00	
4 Reading AIMS	.72	.74	.72	1.00

Correlation Matrix for Fourth Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.72	1.00		
3 Math Benchmark 3	.68	.73	1.00	
4 Math AIMS	.66	.72	.74	1.00

Correlation Matrix for Fourth Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.67	1.00		
3 Reading Benchmark 3	60	.72	1.00	
4 Reading AIMS	.67	.72	.72	1.00

Correlation Matrix for Fifth Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.68	1.00		
3 Math Benchmark 3	.69	.74	1.00	
4 Math AIMS	.68	.71	.78	1.00

Correlation Matrix for Fifth Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.67	1.00		
3 Reading Benchmark 3	.61	.68	1.00	
4 Reading AIMS	.68	.68	.69	1.00

Correlation Matrix for Sixth Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.73	1.00		
3 Math Benchmark 3	.68	.66	1.00	
4 Math AIMS	.69	.64	.70	1.00

Correlation Matrix for Sixth Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.56	1.00		
3 Reading Benchmark 3	.58	.61	1.00	
4 Reading AIMS	.65	.59	.65	1.00

Correlation Matrix for Seventh Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.77	1.00		
3 Math Benchmark 3	.69	.77	1.00	
4 Math AIMS	.72	.77	.73	1.00

Correlation Matrix for Seventh Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.48	1.00		
3 Reading Benchmark 3	.50	.52	1.00	
4 Reading AIMS	.58	.62	.61	1.00

Correlation Matrix for Eighth Grade Math

Test	**1**	**2**	**3**	**4**
1 Math Benchmark 1	1.00			
2 Math Benchmark 2	.68	1.00		
3 Math Benchmark 3	.68	.69	1.00	
4 Math AIMS	.70	.66	.68	1.00

Correlation Matrix for Eighth Grade Reading

Test	**1**	**2**	**3**	**4**
1 Reading Benchmark 1	1.00			
2 Reading Benchmark 2	.56	1.00		
3 Reading Benchmark 3	.45	.48	1.00	
4 Reading AIMS	.63	.66	.55	1.00

Determining whether or not students have mastered standards is a categorical decision. Consequently, the validity of a benchmark assessment program is informed not only by the correlations among benchmark tests and statewide tests, but also by the accuracy of forecasted state classifications of standards mastery. Table 4 shows the accuracy of forecasted classifications for the same district used to illustrate validity evidence involving correlations among benchmark tests and the statewide AIMS assessment. The table shows the percentages of students correctly forecasted from district benchmark assessments to meet or not meet state standards as measured by their AIMS performance. Three forecasts are shown in the table. The first is the forecast from Benchmark 1 to AIMS. The second is the forecast involving the combined mastery patterns in Benchmarks 1 and 2 and AIMS, and the third is the forecast including the combined mastery patterns for all three benchmark assessments and AIMS. The lowest levels of accuracy for the three forecasts were 84, 85, and 85 percent respectively. The highest accuracy levels were 93, 94, and 94 percent. The means were 89, 90, and 90 percent.

TABLE 4
Percent of Students Whose State Mastery Classification Was Correctly Forecasted

		Benchmark Forecasts		
Grade	**Subject**	**1**	**1 & 2**	**1,2 & 3**
Third	Math	92	93	93
	Reading	90	92	92
Fourth	Math	91	93	93
	Reading	90	92	92
Fifth	Math	90	91	93
	Reading	87	88	90
Sixth	Math	90	91	91
	Reading	87	87	88
Seventh	Math	93	94	94
	Reading	84	86	85
Eighth	Math	85	85	86
	Reading	86	88	88

Before closing the discussion of validity, it is worth noting that the illustrative data shown here revealed variability both in the magnitude of reported correlations and in the percentages of correctly forecasted mastery classifications. Of course, the presence of variability is to be expected. That expectation indirectly hints at the importance of validating district benchmark assessment initiatives locally and continuously. Benchmark assessments are designed to articulate to district curriculums, which vary from grade to grade and from district to district. Moreover, curriculums and statewide assessments change continuously. For these reasons, continuous validity studies play an important role in insuring the ongoing effectiveness of benchmark assessment initiatives.

II. Scores for Benchmark Assessments

Several types of scores are available in Galileo reports. These scores fall into three categories. One category is based on the view that a test score is simply a summary of performance on the items on the test in question. No inferences beyond performance on the test are explicitly implied for this type of score. Scores in this category may be referred to simply as observed test scores because they imply no inferences beyond observed test performance. The second category reflects the assumption that test performance should be scored by referencing the level of performance to a pre-established criterion. Scores within this category may be referred to as criterion-referenced scores since they are based on assumptions underlying criterion-referenced assessment (Glaser, 1961) and since they have been widely used in criterion-referenced assessment initiatives. The third score category is based on the assumption that student responses to test items are indicators of their level of performance related to an underlying level of proficiency or ability (e.g., Thissen & Wainer, 2001). Scores falling in this category are estimates of the ability that the test is designed to assess. For example, student performance on a fifth grade math test would provide an estimate of fifth grade math ability. Scores providing ability estimates can be referred to as ability scores.

Observed test scores are widely used in classroom formative assessments designed to assess specific skills that have been targeted for instruction. These scores are often used in grading classroom tests, and they may play a role in determining report card grades. Criterion-referenced scores are widely used to assess the mastery of instructional objectives. In Galileo, criterion-referenced test scores are used to assess mastery of individual learning standards or performance objectives targeted for instruction in standards-based educational initiatives. Ability scores are provided for benchmark tests in Galileo, which are also used in standards-based education. These scores can be used to plan interventions to promote student learning. For instance, they can be used to estimate the kinds of skills that students are likely to be ready to learn at a given time. They can also be used in measuring student progress within and across years, and they can be used in forecasting performance on other tests of interest including statewide tests used in accountability programs. When the ability score distribution is segmented into categories, ability tests can be used to measure varying levels of standards mastery. The paragraphs that follow outline the types of scores available in Galileo reports and discuss the segmentation of ability scores into categories reflecting varying levels of standards mastery.

A. Observed Test Scores

Two types of observed test scores are included in Galileo reports: the raw score and the percentage score. Both of these types of scores are widely used in school systems

i. Raw Score

The simplest way to summarize performance on a test is to sum student responses to each of the items on the test. This type of score is typically referred to as a raw score or summed score. In the typical case, each item on the test is given one point if it is correct and zero points if it is incorrect. In the one-zero case, the raw score is the number-right score on the test. In some cases, some items may be worth more than one point. For example, a math test might include a series of constructed response items, each of which might be scored from zero to four. The meaning of the raw score is somewhat different for this type of test than for the test containing items scored one or zero. When there are more than two possible scores for an item,

the raw score reflects the number of points earned on the test. The interpretation of the score is assisted by indicating the number of points earned and the number of points that it is possible to earn.

Reports containing raw scores are included in Galileo for use with benchmark and formative tests or with other tests such as end-of-course tests. For example, a teacher might construct a short formative test to assess student learning related to one or two performance objectives. The raw score would provide a summary of student mastery of the items on the test.

ii. Percentage Score

A disadvantage to the raw score is that it provides no way to compare the score attained on one test to the score attained on another test. The percentage score is often used to overcome this disadvantage. In the case in which items are scored one or zero, the percentage score gives the percentage of items responded to correctly on the test. In the case in which items may be worth more than one point, the percentage score gives the percentage of the total available points earned on the test.

The percentage score can be used to place multiple tests on a common scale. When this is done, the tests can be averaged. For example, a teacher may average the percentage scores from a series of classroom tests, and use the averages to determine a grade for each student in the class.

B. Criterion-Referenced Scores

In Galileo Online, criterion-referenced scoring is used for subsets of items related to a particular performance objective or learning standard. For example, consider a benchmark test assessing performance on eight performance objectives, each of which is assessed by five multiple-choice test items. Criterion-referenced scoring would be used to assess level of performance with respect to each of the eight performance objectives. Achievement of particular level of performance would be determined by setting one or more cut points in the scale reflecting possible scores.

In the heyday of criterion-referenced assessment, it was customary to recognize two performance levels associated with a single cut point. For example, suppose that a cut point was set at four out of five items correct. If a particular score, (e.g., two out of five items correct) fell below the cut point, the objective would be classified as not attained. On the other hand, if the score was at or above the cut point (e.g., four items correct) the objective would be classified as attained. Current practice in standards-based education sometimes favors more than two performance levels. For example, a district may identify three performance levels such as falls below the standard, approaches the standard, and meets the standard.

Criterion-referenced scoring supports standards-based instructional initiatives by providing data on student learning that assists teachers to target instruction toward the mastery of specific standards. For example, a teacher who has information regarding student mastery of a series of specific performance objectives is in a good position to plan interventions to promote learning related to those objectives.

C. Ability Scores

There are two types of ability scores, both of which are included in Galileo reports. The oldest type is the norm-referenced score. Norm-referenced ability scores indicate standing in a norm group. They tell how far above or below average a score is in a group comprised of individuals whose performance provides the data used to establish the psychometric properties of the test. Tests built for use in Galileo K-12 Online are designed to provide local norms. Accordingly, the norm group is typically composed of students in a district who took the test. For example, the norm group for a fifth grade math benchmark test would typically comprise the fifth grade students in the district who took the math test.

The second type of ability score is developmental in nature. This type of score indicates the position of a score in a developmental progression of capabilities. The developmental score is particularly useful in standards-based educational initiatives because the score provides information about what students can do currently and what they are likely to be able to learn next. Information of this kind is useful in planning interventions to promote learning.

i. Norm-Referenced Scores

Norm-referenced test scores are generally based on the assumption that ability is normally distributed. The normal distribution is a theoretical distribution specified by a mathematical rule. Parameters applied in the rule include the mean of the distribution and the standard deviation, which reflects the dispersion of scores around the mean. One normal distribution may have a different mean and standard deviation than another. Thus, the normal distribution is really a family of distributions.

No actual distribution of scores can be expected to conform exactly to the normal distribution. However, it is reasonable and useful to assume that actual distributions of ability scores will approximate a normal distribution. Figure 3 shows a normal distribution. Note that the curve in the figure is smooth and continuous and that it extends from minus infinity to plus infinity. Actual distributions of scores do not conform exactly to the smooth curve in the figure, and of course they do not extend from minus infinity to plus infinity. In accordance with standard measurement practice (e.g., Thissen & Wainer, 2001), mathematical procedures based on Item Response Theory are implemented in Galileo K-12 Online to smooth the ability score distribution so that it more closely approximates the normal distribution.

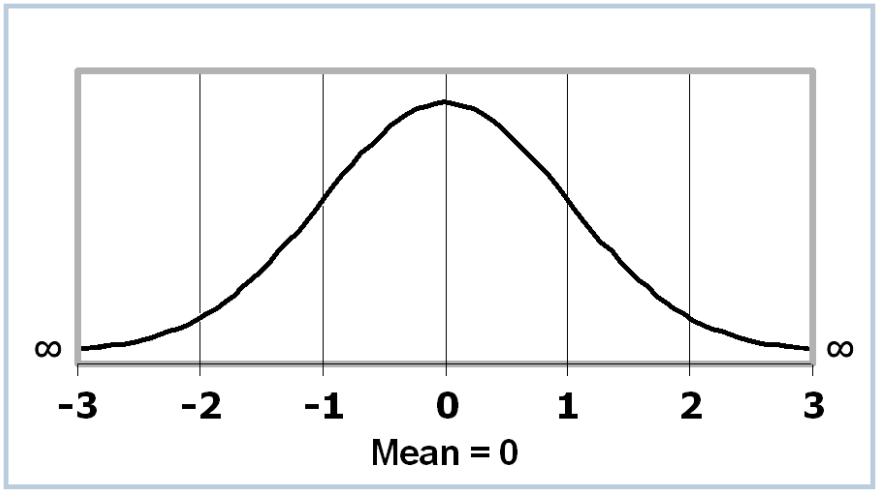


FIGURE 3
Normal Distribution

Four norm-referenced scores are available in Galileo K-12 Online. They are: the Standard Score, the Percentile Rank, the Normal Curve Equivalent Score, and the Developmental Level (DL) Score. All of these scores specify position in a norm group by indicating score position in a normal distribution.

1. The Standard Score - specifies position in a norm group in terms of standard deviation units. A distribution of standard scores has a mean of zero and a standard deviation of one. Figure 3 displays a normal distribution of standard scores. As shown in the figure, a score of plus one is one standard deviation above the mean. A score of zero is exactly at the mean, and a score of minus one is one standard deviation below the mean.

   Standard scores provided in reports in Galileo K-12 Online may be used to plan interventions for individual students or groups of students. For example, a teacher may wish to plan an intervention program for students who score one or more standard deviations below average on a benchmark test.

2. The Percentile Rank - gives the percentage of scores in the norm group at or below a particular score. Figure 4 illustrates the description of position in a norm group in terms of percentile ranks related to a normal distribution. As shown in the graph at the far left in the figure, the 50th percentile falls at the mean of the distribution. Fifty percent of the scores in the distribution fall at or below this point. The middle graph in the figure indicates that the 84th percentile is approximately one standard deviation above the mean. Eighty-four percent of the scores are at or below this point. The graph at the far left in the figure shows that the 16th percentile is approximately one standard deviation below the mean. Sixteen percent of the scores are at or below this point.

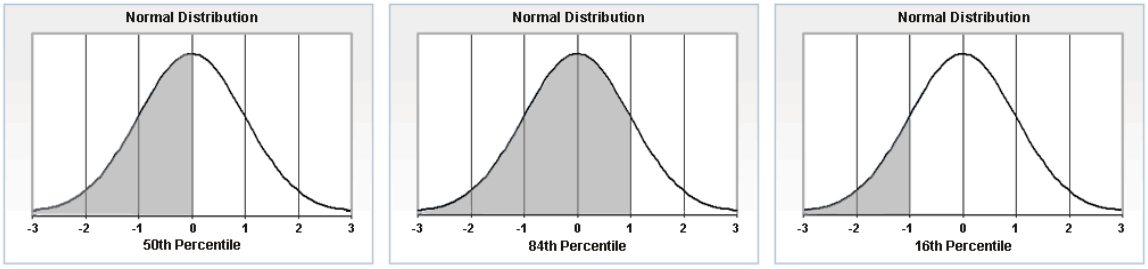


FIGURE 4
Norm Group

Percentile ranks in Galileo reports serve essentially the same purpose as standard scores in that both scores indicate position in a norm group. The percentile rank is particularly useful in communicating to audiences with varying familiarity with statistical terms. Percentages are widely used in a variety of fields to communicate quantitative information. As a result many people have a good idea of what a score expressed in terms of a percentage means.

3. The Normal Curve Equivalent (NCE) Score - is a norm-referenced score with a mean of 50 and a standard deviation of 21.06. The result achieved by setting the mean at 50 and the standard deviation at 21.06 is a set of equal interval scores ranging from zero to 99. NCE scores have been widely used in federally funded remedial education programs. They provide a common approach for describing performance on tests in different subject areas. In so doing they facilitate comparisons of test scores reflecting achievement in different areas. It should be noted that standard scores and percentile ranks also can be used to compare test scores in different areas.

4. The Developmental Level (DL) Score - can be interpreted as a norm-referenced score. However, as its name implies, it can also be interpreted from a developmental perspective. Discussion here will focus on the norm-referenced nature of the DL Score. The DL score is calculated in Galileo K-12 Online using a mathematical model based on Item Response Theory (Thissen & Wainer, 2001). The model assumes that ability is normally distributed and that the ability distribution has a mean of zero and a standard deviation of one. In keeping with accepted measurement practice, a series of linear transformations are implemented to replace the mean of zero and standard deviation of one with numbers that are more intuitively reasonable for teachers, administrators, specialists and parents. The values used for these transformations will vary depending on the extent to which test equating is implemented across grades.

The discussion here illustrates the transformation process when equating is not implemented. The first step is to multiply the standard deviation of the distribution at each grade level by 100. This sets the standard deviation for benchmark tests at each grade level to 100. The next step is to transform the mean at each grade level. At each grade level, the mean of zero is transformed using the formula 500 + (grade level x 100). For example, the mean DL for a fifth grade benchmark assessment would be: 500 + (5 x 100) = 1000. Similarly, the mean DL for a sixth grade benchmark would be: 500 + (6 x 100) = 1100. The mean DL scores for Galileo

benchmark assessments when tests are not equated across grades are provided in the table that follows:

TABLE 5
Means and Standard Deviations for Benchmark Assessments by Grade Level

Grade	Mean DL	Grade	Mean DL
K	500	7	1200
1	600	8	1300
2	700	9	1400
3	800	10	1500
4	900	11	1600
5	1000	12	1700
6	1100		

The norm-referenced interpretation of the DL Score is similar to the interpretation for the standard score. For example, in the fifth grade, a score of 1000 indicates average performance and a score of 1100 indicates performance that is one standard deviation above the average.

ii. Developmental Scores

Norm-referenced scores are useful in identifying which students might profit from a given intervention program, but they are of limited value in determining what should be taught to improve ability. The source of this limitation is that norm-referenced scores do not provide information about the skills that a student possesses or what skills the student is likely to be ready to learn. In the Galileo system, the major purpose for measuring ability is to promote learning. A developmental conception of ability lends itself well to this purpose (Bergan, 1981; Thissen & Wainer, 2001). From the developmental perspective, the ability score indicates a student's position in an ordered progression of capabilities. Knowing the student's position in the developmental continuum makes it possible to anticipate the kinds of skills that the student will be capable of learning as development progresses.

In Galileo K-12 Online, the DL score is used to estimate position in a developmental progression. As indicated earlier, the DL score is computed using mathematical algorithms based on IRT (Thissen & Wainer, 2001). IRT models are particularly well suited to a developmental approach to assessment because they place ability and item difficulty on the same scale. For example, a DL score of one standard deviation above the mean corresponds to an item difficulty of plus one. Likewise, a DL score of one standard deviation below the mean corresponds to an item difficulty of minus one. IRT models make use of the correspondence between ability and item difficulty to estimate the likelihood that students with a given DL will be able to perform items of varying difficulties. This information may be used to target specific skills for instruction. For example, suppose that a teacher is planning an intervention for a small group of students whose DL scores are not high enough to meet a given standard. The DL scores can be used to identify specific skills that may be targeted for instruction and that, if mastered, will lead to DL scores that do meet the standard.

D. Benchmark Score Categories Reflecting Standards Mastery

In standards-based educational initiatives, there are circumstances in which discrete levels of performance are constructed by breaking a continuous score distribution into score categories. For example, scores on statewide assessments are broken into discrete categories reflecting varying levels of standards mastery. Various procedures (see, for example Cizek, 2001) may be used to determine the cut points for levels bearing labels such as "approaches the standard" or "meets the standard." All of these procedures involve an element of subjective judgment (Cizek, 2001).

Score categories are useful for defining instructional goals. For instance, if students performing at or above a particular score are classified as meeting a certain standard, then it is reasonable to set meeting that standard as an instructional goal for students. Achievement of an instructional goal defined by score categories implies the mastery of interrelated sets of skills. A major benefit to the score category approach to the specification of instructional goals is that it provides guidance for large blocks of instruction covering multiple skills that are related to each other.

When cut points have been established for score categories defined at the state level, it is useful to construct score categories for benchmark assessments that are aligned to the state cut points. Aligning benchmark cut points to state cut points increases the likelihood that levels of standards mastery determined from local benchmark assessments are linked to levels of mastery established at the state level. By contrast, when cut points are not aligned, linkage is not likely to be adequate. For example, if 60 percent of the students in a particular district were classified as meeting the standard based on statewide test performance and 90 percent of the students were classified as meeting a benchmark test standard, then the benchmark standard would very likely overestimate the number of students classified by the state as meeting the standard.

Alignment of benchmark cut points to state cut points can be achieved in a number of ways. Equipercentile equating is the preferred approach used in Galileo Online for aligning benchmark cut points to statewide test cut points. In this approach the cut point on the benchmark test is the score at the percentile rank corresponding to the percentage of students meeting the standard based on statewide test performance. In those instances in which information on statewide test performance is not available, default benchmark goals are set. These goals reflect reasonable expectations for student learning. However, they are ***not*** linked to statewide test score categories.

Alignment of benchmark cut points to cut points established for statewide tests is very useful. However, ***alignment does not imply that a student who fails to meet benchmark goals will necessarily fail to meet standards based on statewide test performance. Nor does it suggest that a student who meets benchmark goals will necessarily meet standards based on statewide test performance. Forecasts can never be expected to be accurate 100 percent of the time.***

When benchmark goals are aligned with state standards, the ability to accurately forecast the ***number*** of students likely to meet state standards is often markedly enhanced. However, alignment is only one issue that needs to be addressed in a forecasting initiative. The accuracy of benchmark forecasts regarding ***which*** students are likely to meet state standards based on benchmark test performance is affected by several factors in addition to benchmark alignment. These factors include the reliability of the benchmark instruments, benchmark validity

assessed by the magnitude of the relationship between benchmark tests and the statewide test, changes in standards mastery cut points instituted at the state level, changes in the statewide test, and changes in the benchmark tests. Given the number of variables that may affect forecasting precision, variations in forecasting accuracy are to be expected. For example, in a recent study of benchmark forecasting in multiple grades in four school districts, accuracy in forecasted classifications ranged from 73 percent to 94 percent with a mean of 87 percent.

III. Reports for Benchmark Assessments

Galileo Online contains a broad range of reports that can be used with varying types of assessments. The focus here is on four reports that are particularly useful for guiding instruction based on benchmark assessment results. The first report that will be discussed is the *Item Analysis Report*. This report is useful for planning interventions at the item level. For example, the report reveals the kinds of distractors that students choose. This information can be used to identify typical errors in thinking. Interventions can then be designed to address those errors. The *Item Analysis Report* also provides information that can be used to improve student test-taking skills. The second report to be addressed is the *Development Profile*. This report can be used to identify the performance objectives that students have mastered and those that have not been mastered. Results given in the *Development Profile* can be used to plan interventions targeted at specific objectives for specific groups of students. The third report is the *Development Summary*. This report provides information on the relative standing of students in a local norm group. The report is useful in identifying individual students and groups of students that may benefit from one or more interventions. The fourth report is the *Aggregate Multi-Test Report*. This report provides information on performance on multiple tests. When information on statewide test performance is available, the report can be used to identify students who have met benchmark standards corresponding to state standards. The report is especially useful for identifying students who are on course to meet state standards and those who are at risk for not meeting standards. The report is also useful for identifying the specific objectives that students need to master to increase their likelihood of meeting state standards. This information is very useful in planning specific interventions.

A. Item Analysis Report

The *Item Analysis Report* provides information about the responses students have made on benchmark assessment. It can be used to evaluate the numbers of students who answered a given question correctly as well as the counts of students who selected each distracter. It can also be used to evaluate the responses broken down by overall score on the assessment.

i. Report Information

The *Item Analysis Report* displays the total number of students who selected each available option on the items making up a benchmark assessment. The detailed item analysis format provides this data broken down by the student's percentile score on the overall assessment. Clicking on the aggregate data in any box yields a list of those students whose responses are summarized. Clicking on the item number provides the user with the item text in a popup window. The examples that follow show the data that is displayed.

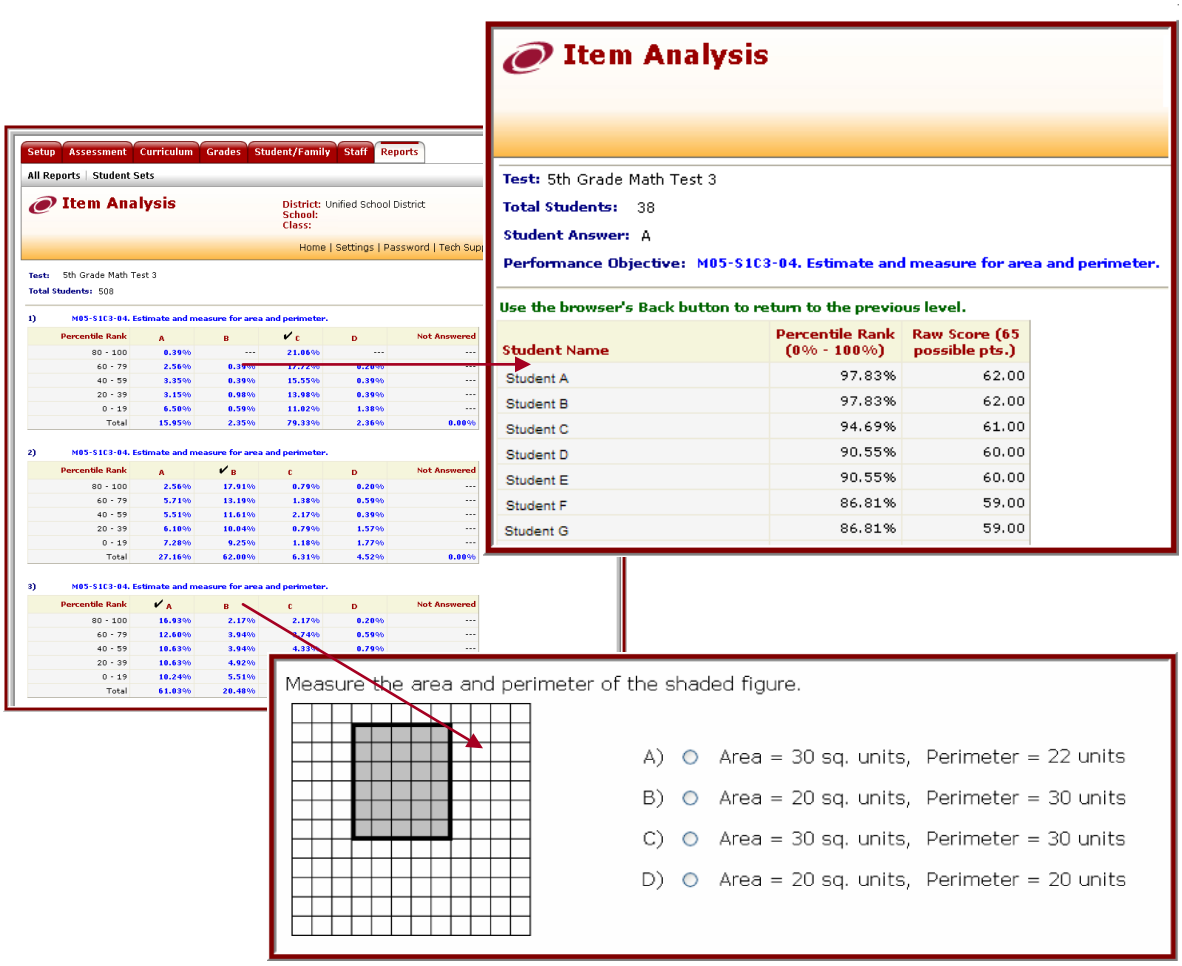


FIGURE 5
Item Analysis Report with question and student percentage ranking drill downs.

ii. Administrative Report Uses

Administrators will usually run this report aggregated at the school or district level. It provides a measure of student mastery of items related to objectives that may be targeted for instruction. It also provides an indication of the types of errors being made by students. In the case of the detailed item analysis, the data indicate how those students who did well on the assessment responded to a given item as compared to those who didn't achieve high scores. The information could be used to guide decision-making for the development and implementation of intervention programs. If large numbers of students are making particular types of errors, then that information could serve as a guide for what should be covered in an intervention program. If large numbers of students, both high and low scoring alike, are missing items covering a particular skill, then an administrator may wish to explore whether that performance objective is receiving adequate coverage in the curriculum materials. This question may be explored through the use of a curriculum map.

iii. Classroom Intervention Uses

The classroom uses for the *Item Analysis Report* are similar to those of the administrator uses in that in both cases the interest is in finding out which students are able to answer the questions correctly as well as learning what types of errors that are being made. In the case of the teacher, the report would be run at the class level. The results may provide a very different

overall picture than the district results. Whereas a principal might find that large numbers of students are not demonstrating mastery on a particular math performance objective, a given math teacher might not have similar findings. Their planning for material to be covered in class or possible interventions that would be required would be shaped accordingly.

B. Development Profile Report

Galileo Online includes reports that specify varying levels of mastery for performance objectives assessed with multiple items. The system can indicate those performance objectives that the student will probably be ready to learn now or in the near future, and those capabilities that the student will be ready to learn later. The *Development Profile Report* specifically indicates whether or not performance objectives have been achieved and the mastery levels for each performance objective that has been assessed, but not mastered.

i. Report Information

The *Development Profile Report* displays the performance objectives covered by a benchmark assessment. The percent of students achieving each of the available mastery levels are shown beside the performance objective. The report can be run in this format at the district, school, and class levels. When run at the individual level, the mastery level achieved by the student is shown. The screen shots below show examples of both the aggregate and individual formats for the *Development Profile*.

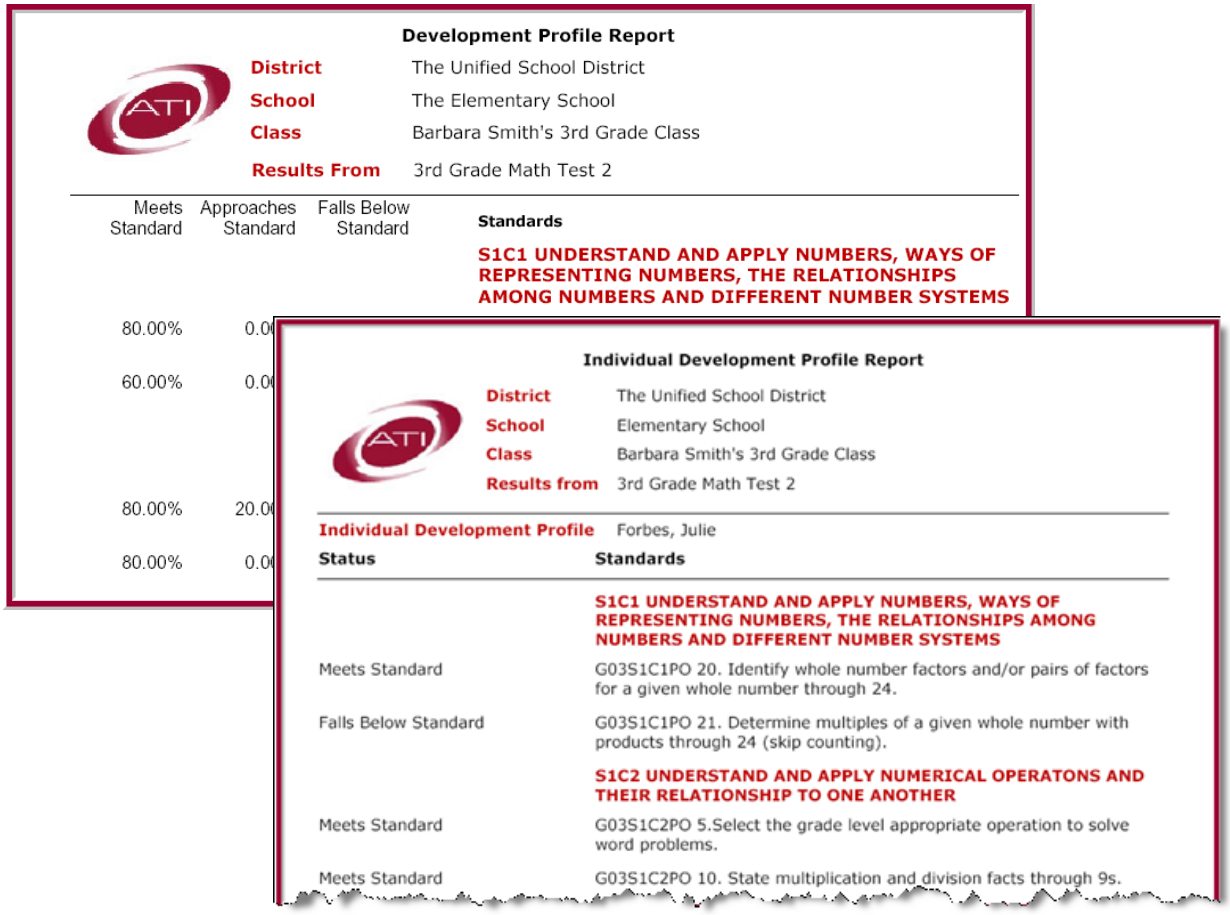


FIGURE 6
Development Profile Report

ii. Administrative Uses

The *Development Profile* provides the administrator with an efficient means of tracking student mastery of different performance objectives at many different levels of aggregation. The report can be run at the district, school, or class level. It can also be run for groups of student that are defined by various demographic variables such as ethnicity or ELL status.

Examination of the results for individual performance objectives could guide the administrator to examine whether individual performance objectives are being adequately covered in the curriculum or to evaluate whether an intervention program is having the desired effect. In the first instance, if the results show that students have not mastered a given performance objective at the desired level, then a curriculum map could be run to determine if those objectives are being adequately covered in the curriculum. When planning interventions, the mastery level data would provide very specific information that could be used to identify the students who would likely benefit from the intervention as well as identifying the specific performance objectives that should be the focus.

iii. Classroom Intervention Uses

The uses of the *Development Profile Report* for the teacher are much the same as they would be for the administrator. Both would use the information to evaluate mastery of standards and

to make plans accordingly. In the case of the teacher, the data might be used to identify those performance objectives that might warrant additional coverage in class. The teacher might discover that a performance objective had not been mastered at the level hoped even though it had been covered in the classroom. The data might also be used to evaluate whether there were specific subgroups of students that might profit from a focused intervention. In many cases, the results obtained at the classroom level may be quite different from the results indicated by aggregated reports being run at the district or school levels.

C. The Development Summary Report

The *Development Summary Report* supplies a series of scores that indicate position in a norm group. Typically, local norms are used for the report. The various scores offer options useful in communicating results to different audiences including students, parents, educators, and others with a special interest in student achievement.

i. Report Information

The *Development Summary Report* provides four scores: The Percentile Rank, Standard Score, Normal Curve Equivalent Score, and Developmental Level Score. All of these scores can be interpreted as norm-referenced scores and each is discussed previously in Section II.

The norm group for a given test is the group of students comprising the standardization sample for the test. In Galileo K-12 Online reports, this sample is typically composed of the students in the district who took the test. For example, the standardization sample for a fifth grade math benchmark test would typically be comprised of the fifth grade students in the district who took the test.

Each of the four scores in the *Development Summary Report* indicates the standing of a student or group of students in the norm group. For example, the percentile rank for a class indicates the average percentile for that class. If the percentile is well above 50, the class on average has scored well above the 50^{th} percentile.

When the *Development Summary Report* is run at a district level, it provides the four scores described above for each school in the district. A drill down feature shows the scores for each class in a school. An additional drill down shows the scores for individual students. If the report is run at the school level, scores for all the classes in the school at a given grade level are shown. If the user drills down on a given class, the scores for all the students in the class are shown.

ii. Administrative Report Uses

In standards-based initiatives, norm-referenced test scores are typically used to identify individuals or groups that deviate substantially from the norm-group mean. For example, if the average score for a particular school were more than a standard deviation below the mean for the district, that school might be targeted for a special intervention program.

The *Development Summary* provides a quick and convenient way to identify groups of students that may profit from various intervention initiatives. For example, an administrator might target classes of students with scores well below average for a reteaching intervention.

iii. Classroom Interventions

Teachers can use the *Development Summary Report* to identify individual students who may benefit from reteaching or enrichment activities. For example, a teacher may wish to provide additional learning opportunities for those students scoring well above average on a benchmark assessment.

D. Aggregate Multi-Test Report

The *Aggregate Multi-Test Report* provides information from multiple assessments. It can be used to assess student progress both within and across years. It can be used to compare performance on benchmark assessments to performance on external tests such as statewide assessments. It can be used to assess student risk of not meeting standards and it can be used to target specific learning objectives leading toward standards mastery.

i. Report Information

The *Aggregate Multi-Test Report* is capable of displaying results from one or more external tests or internal tests such as benchmark tests. When benchmark tests are selected, the report provides a *Developmental Level Score* for the selected tests. In the case in which two or more tests are equated, progress can be assessed. For example, if the score on the first benchmark test was 1000 and the score on the second was 1150, the user would know that substantial progress occurred from the first benchmark assessment to the second benchmark assessment.

If statewide test scores are available from a previously administered statewide test, cut points for standards mastery can be set to correspond to cut points on the statewide test using the equipercentile equating procedure described earlier. For example, if a given cut point for the statewide test was at the sixtieth percentile, a corresponding cut point could be set at the sixtieth percentile for the benchmark test.

A drill-down feature identifies average scores for groups of students as well as the mastery classification associated with each score. For example, suppose that the average score for a class was 950, the corresponding mastery classification might be “Approaches the Standard.” The mastery classification categories define goals for benchmark test performance. For example, a goal might be to achieve a score indicating that standards have been met.

The *Aggregate Multi-Test Report* allows the user to drill-down to the point at which individual students are displayed and their mastery of performance objectives covered in a benchmark test is indicated. For example, the report might reveal that a particular student approached the standard on a given set of objectives and that the student had met the standard with respect to another set of objectives. The report would also indicate the specific objectives to be targeted for instruction to promote standards mastery. For instance, suppose that a student received a benchmark test score with a mastery classification of “Approaches the Standard.” Suppose further that the report targeted four objectives for instruction and that the student subsequently mastered those objectives. If a new benchmark test score reflected what the student had previously learned and mastery of the targeted objectives, the new benchmark score would indicate that the student had met the standard for the benchmark assessment.

When forecasting information is available from a previous year, risk levels can be assigned based on patterns of standards mastery across multiple benchmark tests. For

example, students failing to master standards on successive benchmark tests may be classified as highly at risk for failing to master standards assessed through performance on statewide tests. Information on the accuracy of previous forecasts can also be provided.

The *Aggregate Multi-Test Report* can display scores on one or more external tests as well as scores on benchmark tests. For example, the *Multi-Test Report* can display scores on statewide tests. Moreover, external test scores may be displayed along with benchmark test scores.

ii. Administrative Report Uses

The information provided by *Multi-Test Report* can be used to plan and allocate resources for interventions aimed at promoting student learning.

- DL scores for groups of students coded for standards mastery tell the administrator which school or classes have average scores falling below standards mastery cut points. This information can be used to target intervention resources toward schools or classes failing to meet standards.

- Information targeting specific objectives for instruction tells the administrator what objectives should receive attention in intervention plans. This information helps the administrator to promote the design of interventions that are likely to be maximally effective in promoting standards mastery.

- Patterns of mastery across multiple tests can indicate groups of students at varying levels of risk for failing to meet standards. This information can be used to target interventions toward students at risk for not mastering standards. For example, short term intensive interventions may be planned for students at high levels of risk late in the school year.

- For equated tests, information on progress from one assessment to the next can indicate the extent to which classes or schools are showing gains during the year. Administrators can use progress information to target interventions toward schools or classes failing to show progress.

- Information comparing performance on a series of external tests (e.g., statewide tests) allows the administrator to examine trends in statewide test performance over time. This information can be used to target intervention resources. For example, in classes where statewide test performance is declining, intervention may be warranted.

iii. Interventions and the Aggregate Multi-Test Report

Classroom teachers and other individuals and groups engaged in interventions to promote student learning can use the *Multi-Test Report* to design interventions for individual students as well as for groups of students with common instructional needs. In some cases, interventions may occur in the context of regular classroom instruction. In other instances, interventions may require instruction in addition to that occurring in the classroom. In some instances, interventions may involve emersion in instructional content over an extended time span. In other cases, interventions may involve intense exposure to a limited number of objectives during a short time period.

- The drill-down feature in the *Multi-Test Report* that identifies specific objectives to be targeted for instruction with specific students provides the information necessary to plan individualized intervention programs. For example, if a *Multi-Test Report* indicated that instruction should focus on three objectives for a small group of students, an intervention plan could be customized to address those objectives.

- When tests are equated, report information on progress can be used to alert teachers early on to the need for interventions involving individual students or small groups of students. For example, if a student falls behind from one benchmark test to the next, an intervention may be warranted.

- Information on patterns of standards mastery across multiple benchmark tests can be used to identify students who are highly at risk for not meeting standards. For example, if a group of students were to fail three consecutive benchmark tests, they might be classified as highly at risk for not meeting standards. An intensive intervention program could be designed to address the learning needs of these students.

IV. Meeting Educational Challenges in the 21st Century

Benchmark assessment and standards-based education are new concepts developed to meet the challenges of a new age. They are two among many innovations designed to promote learning necessary for the continuing development of an effective citizenry capable of shaping the global community in which we all live in beneficial ways. In a world that changes as rapidly as the one in which we now live, it is reasonable to assume that the concepts of benchmark assessment and standards-based education will themselves change quickly. Yet, the need for instruction that is goal directed and educational management that adapts in effective ways to maximize the likelihood that goals are met will not change. What we can expect and, indeed, what we are currently witnessing is a rapid evolution of human commitment and innovative technology designed to support goal-directed, data-driven instruction. Benchmark assessment and standards-based education are a part of this evolution.

V. References

Bergan, John R. (1981). Path referenced assessment in school psychology. In Thomas R. Kratochwill (Ed.), *Advances in school psychology* (2nd ed., pp. 255-280). Hillsdale, NJ: Erlbaum.

Bergan, Bergan, & Guerrera, (2005). *Standards Mastery Determined by Benchmark and Statewide Test Performance.* Tucson, AZ: Assessment Technology Inc.

Cizek, G.J. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Clogg, C.C., & Eliason, S.R. (1987). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working Paper 1977-09). University Park, PA: Population Issues Research Office.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 18, 519-521.

Goodman, L.A. (1974a). The analysis of systems of quantitative variables when some of the variables are unobservable: Part I, A modified latent structure approach. *American Journal of Sociology, 79,* 179-259.

National Research Council (1999). *High Stakes: Testing for Tracking, Promotion and Graduation.* Washington, D.C.: National Academy Press.

Thissen, D. & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, N.J.: Lawrence Erlbaum Associates.

This page intentionally left blank.