Building Reliable and Valid Benchmark Assessments

A Resource for District/Charter School Leaders and Staff Responsible for Benchmark Planning, Construction, and Review

by
John Richard Bergan, Ph.D.
John Robert Bergan, Ph.D.
Kathryn S. Bergan, Ph.D.
Christine G. Burnham, Ph.D.
Scott A. Cunningham, B.S.
Jason K. Feld, Ph.D.
Karyn L. White, M.A.,
and Kerridan A. Kawecki, B.A.

ATI™
Assessment
Technology
Incorporated

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

Assessment Technology, Incorporated, Publishers
Tucson, Arizona, U.S.A.
Printed in the United States of America.
V9-060313

Building Reliable and Valid Benchmark Assessments
A Resource for District/Charter School Leaders and Staff Responsible for Benchmark Planning, Construction, and Review

By John Richard Bergan, Ph.D.; John Robert Bergan, Ph.D.; Kathryn S. Bergan, Ph.D.;
Christine Guerrera Burnham, Ph.D.; Scott A. Cunningham, B.S.;
Jason K. Feld, Ph.D.; Karyn L. White, M.A.; and Kerridan A. Kawecki, B.A.
Assessment Technology, Incorporated

Table of Contents

I. Introduction

As a part of Assessment Technology Incorporated's (ATI) partnership with districts and charter schools, ATI provides district/charter school staff with guidance and recommendations to assist in the planning, construction, and review of a variety of types of customized assessments including benchmark assessments, pretests, posttests, placement, screening, and end-of-course assessments. This resource document is designed for use by district/charter school leadership staff responsible for oversight and supervision of assessment planning, construction, and review activities. Although this document focuses on ATI guidance related to the creation of valid, reliable benchmark assessments, many of the recommendations presented here also apply to the creation of other types of assessments. ATI Educational Management Services staff will provide districts/charter schools with additional guidance related to the creation of other types of assessments as needed.

Throughout the year, this document will serve as an important resource as ATI Educational Management Services staff and district/charter school staff work together to create customized benchmark assessments that accomplish district/charter school goals including:

- Demonstrating alignment to state and district standards including Common Core State Standards, as well as statewide test blueprints (e.g., Arizona's Instrument to Measure Standards [AIMS], Partnership for Assessment of Readiness for College and Careers [PARCC]), Smarter Balanced Assessment Consortium [SBAC], and district/charter school pacing guides.
- Providing documentation throughout the school year on student progress towards standards mastery for teachers, administrators, specialists, and parents to utilize in standards-based planning, instruction, and intervention.
- Providing reliable and valid information on student achievement and growth throughout the year that can be used as part of district/charter school initiatives related to instructional effectiveness to guide professional development, student intervention, and other activities.
- Providing periodic forecasts of student risk for failing to meet standards as measured by statewide tests.

II. Overview of Benchmark Assessments

A. Benchmark Assessment Characteristics and Uses

i. Standards-Based Assessment to Inform Instruction

Benchmark assessments are locally relevant, district-wide assessments designed to measure student achievement of standards, including Common Core State Standards, for the primary purpose of providing information to guide instruction. The first step in constructing a benchmark assessment is to identify the standards that the assessment will measure. Galileo® K-12 Online benchmark assessments can measure student mastery of standards targeted for instruction. In so doing, they indicate what students have accomplished when given appropriate learning opportunities. Benchmark assessments also inform instruction in cases where standards have not been mastered even though appropriate learning opportunities have been provided.

Benchmark assessments provide information to guide instruction in a 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 assessment designed to assess mastery of standards covered during the initial instructional period. For example, a benchmark assessment may be administered after instruction implementing a set of unit plans has been completed. Teachers and administrators may use the results of that assessment to plan and implement interventions to address areas in which students may not have displayed mastery of standards measured on the test. For instance, a reteaching intervention may be employed, assisting students in mastering standards that they have not yet met. Short formative assessments built using *Test Builder* or accompanying interventions administered through ATI *Instructional Dialogs* may then be used to ensure that standards not initially mastered have been mastered following reteaching.

ii. Providing Multiple Assessments of Standards Mastery

Typically the cycle of teaching, assessment, and intervention with benchmark assessments supported by short formative assessments is 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. The multi-test approach has several benefits for the student and district/charter school stakeholders:

- Any test, including statewide achievement tests used to make high-stakes decisions, has some degree of measurement error, use of the multi-test approach can reduce the impact of a single test on high-stakes decisions.
- The multi-test approach increases the likelihood that the assessments used to assess mastery cover the full range of content that has actually been taught.
- The multi-test approach increases timely access to assessment information that can be used in the overall determination of mastery.
- Multi-test information can be used in providing evidence of standards mastery in cases in which other available evidence is insufficient or subject to question.

B. Benchmark Assessment Psychometrics

ATI routinely conducts psychometric analyses of district-wide assessments such as benchmark assessments. A variety of methods including procedures based in Item Response Theory (IRT) are used to estimate item parameters and student ability, to place assessment scores on a common scale to evaluate growth, and to evaluate the reliability, validity, and forecasting accuracy of assessments. District and charter school staff involved in the creation of benchmark assessments will find it helpful to have a general understanding of these approaches as well as a familiarity with ATI recommendations related to the design of valid, reliable assessments that accurately forecast statewide test performance.

i. Estimating Item Parameters and Student Ability

An essential component of the psychometric analysis of all Galileo® benchmark assessments is establishing item parameter estimates using 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 that includes three item parameters: A discrimination parameter, a difficulty parameter, and a guessing parameter. The three parameters are described in the following.

1. Discrimination Parameter

The 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 zero discriminate between different ability levels very poorly. The discrimination parameter indicates the relationship of the item to the underlying ability being measured divided by measurement error. Items with high discrimination values are desirable for inclusion on assessments as they provide a positive contribution to test reliability.

2. Difficulty Parameter

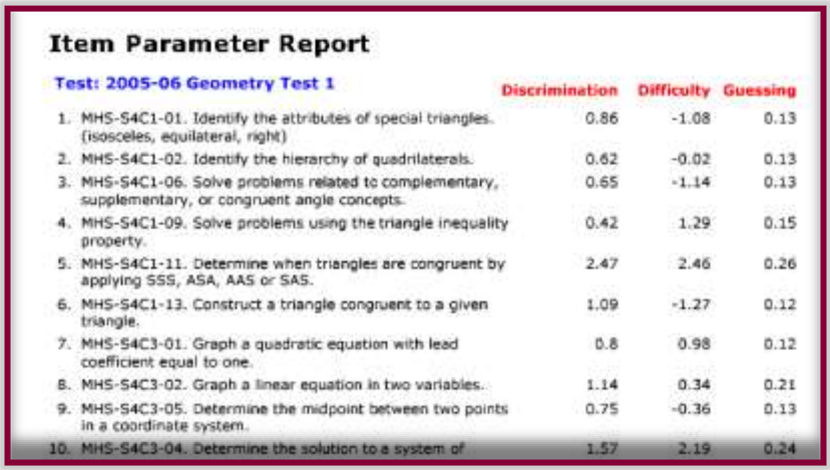
The difficulty parameter provides information on the relative difficulty of an item. An item with a difficulty parameter of zero is at about an average level of difficulty. If the item difficulty is above zero, the item is of above average difficulty. When the item difficulty is negative, the item is of below average difficulty. In general, it is useful to construct benchmark assessments that include a broad range of difficulty levels. Benchmarks of this kind will be sensitive to a range of ability levels and generally correlate higher with criterion measures (e.g., statewide assessments) than benchmarks sensitive to a limited range of ability levels.

3. Guessing Parameter

The guessing parameter indicates the likelihood a student who does not know the answer to 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 Galileo® benchmark assessments is provided through the *Item Parameter Report* available in Galileo K-12 Online. Below is a sample *Item Parameter Report.* During the test review process described later in this document, the final reviewer can also see all available established item bank parameters for items on the assessment.

Item Parameter Report – ***Confidential Screen Shot***



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	1.57	2.19	0.24

ii. Placing Assessment Scores on a Common Scale to Evaluate Growth

ATI uses advanced psychometric techniques based on IRT to place test scores from district-wide assessments (e.g., benchmark assessments) on a common scale to measure student growth. This scaling makes it possible to measure progress without repeating large amounts of assessment content across assessments. The details of the scaling method used by ATI are described in the Galileo® K-12 Online technical manual, available at http://www.ati-online.com/pdfs/researchK12/K12TechManual.pdf. A brief summary follows.

There are a number of ways to place scores from multiple assessments on a common scale. The process used by ATI utilizes IRT procedures and directly links the task of placing scores on a common scale to the estimation of item parameters. When a district/charter school completes the administration of a benchmark assessment, ATI analyzes the student data using IRT and established item parameters to estimate student ability. Item difficulty and item discrimination parameters are involved in placing ability scores on a common scale. These parameters are directly associated with the determination of the mean and standard deviation of the ability distribution.

Scaling for two assessments is accomplished by the anchor item approach. Each assessment includes a set of items from the Galileo secure assessment banks that are on a common scale. When each IRT analysis is run, the parameter estimates for the items in the assessment are fixed to the established, common-scale parameter estimates. In this way, the ability estimates for the students are pulled onto the common scale. This scaling process ensures that student Developmental Level (DL) scores for all district-wide assessments within a grade and content area end up on the same, common scale, and so student DL scores across district-wide assessments are comparable and can be used to measure growth during the school year. During the creation of benchmark assessments, ATI Educational Management Services staff will help ensure that the assessment has sufficient items with established item bank parameters (typically 30-50 percent) to support the scaling process. ***During the test review process, final reviewers should be aware that replacing a large number of items with established parameters with items without parameters may impact the scaling process.*** The final reviewer can see all available established item bank parameters for items on the assessment.

iii. Evaluating Reliability, Validity, and Forecasting Accuracy

ATI conducts ongoing research on the reliability, validity, and forecasting accuracy of a variety of types of district-wide assessments such as benchmark assessments. The results of these analyses routinely indicate that Galileo® assessments are reliable, are valid, and are effective forecasters of student performance on statewide assessments.

1. Benchmark Reliability

Benchmark assessments 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 is internal consistency. Measures of internal consistency provide information regarding the extent to which all of the items on a benchmark assessment are related to the underlying ability (Developmental Level) that the assessment is designed to measure. Benchmark assessments such as those provided to districts and charter schools through Galileo K-12 Online are designed to correlate with other measures of student proficiency including statewide assessments. An assessment that lacks internal consistency does not correlate well even with itself. Therefore, it is unlikely that it would correlate well with other measures.

Reliability is directly affected by the length of the assessment. Longer assessments tend to be more reliable than shorter assessments. The illustration that follows plots the relationship between test length and reliability for 2,018 Galileo benchmark assessments in math, reading/English language arts, science, and writing for grades kindergarten through 12. The data in the figure suggest that test reliability begins to stabilize at above 0.80 for assessments with approximately 50 items or more. ATI recommends that benchmark assessments contain a *minimum of 45 items* to help ensure adequate reliability.

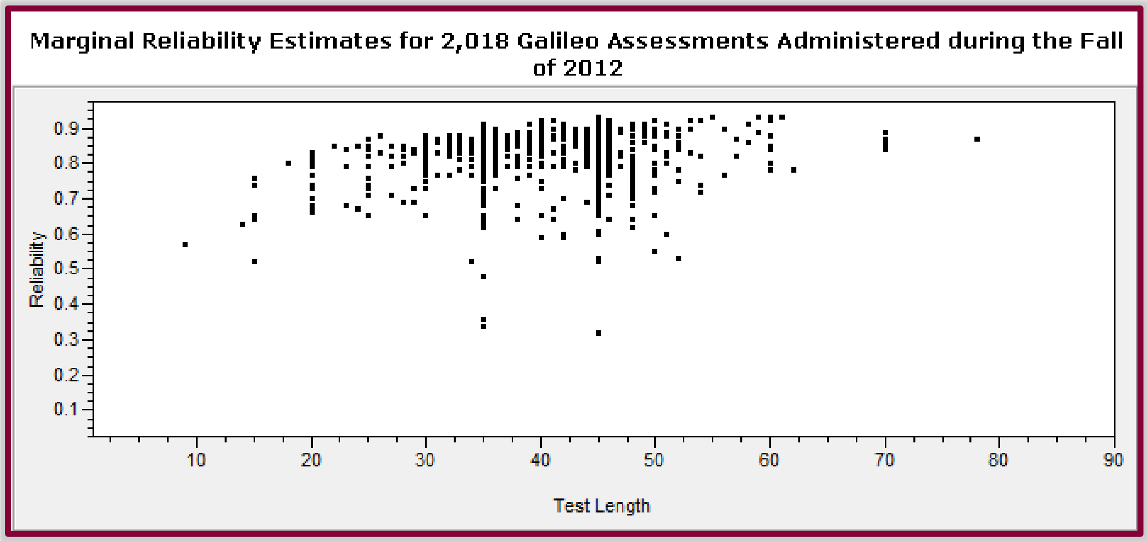


Figure 1
Test reliability of 2,018 benchmark assessments as a function of test length

2. Benchmark Validity

Standards-based educational initiatives across the nation are targeting instruction toward the achievement of local and state standards including Common Core State Standards. State standards provide a common set of goals for all districts and charter schools 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 predictive validity of the benchmark assessments. Although significant correlations do 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 and local education agencies and for 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.

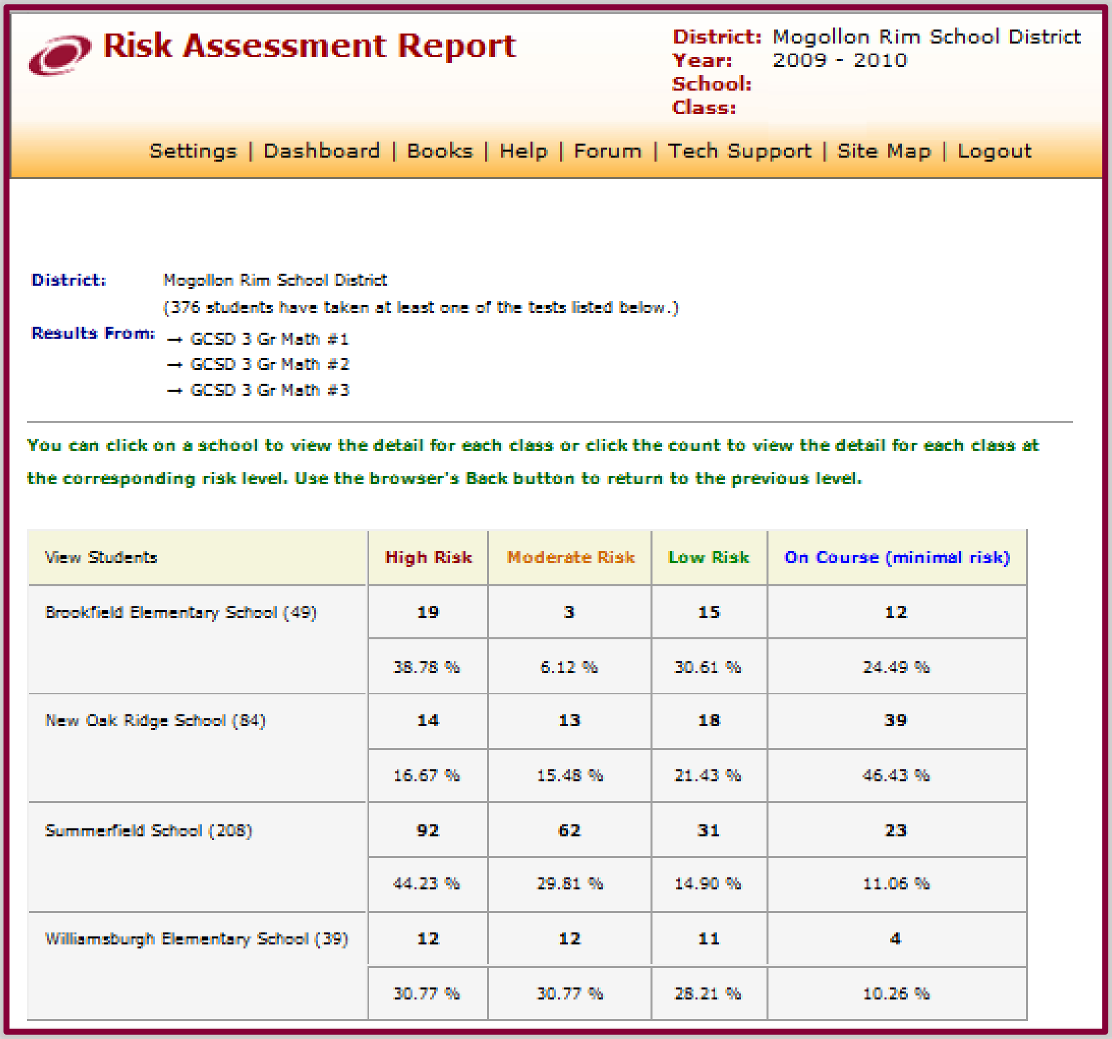
ATI maintains a continuous research program to establish the predictive validity of district-wide assessments (e.g., benchmark assessments) as evidenced by the correlation between each district-wide assessment and the statewide assessment. Investigations of predictive validity are carried out on an annual basis once districts/charter schools have uploaded their statewide assessment data for individual students. The correlation coefficients that result from these investigations are presented as part of the Galileo® *Forecast Report* which will be described later in this section. The continuous examination of predictive validity makes it possible to track variations in validity that may be associated with variations in the content of assessments, variations in the statewide test, and variations in instruction, such as those associated with the introduction of Common Core State Standards and new statewide assessment programs (e.g., PARCC, Smarter Balanced Assessment Consortium [SBAC]).

The ongoing predictive validity research that ATI conducts consistently indicates that student performance on Galileo district-wide assessments is correlated with student performance on the statewide assessment. In a recent investigation ATI evaluated the correlations for 1,185 Galileo assessments administered in the 2011-12 school year by the first 25 districts/charter schools located in Arizona, California, Colorado, and Massachusetts who submitted their spring 2012 statewide assessment data. The average correlation coefficient between the Galileo assessments and the statewide tests was 0.75, which is very consistent with what has been observed in past years. The average correlation coefficients for reading, math, and science assessments are 0.74, 0.75 and 0.73 respectively. These strong correlation coefficients indicate that the scores students achieve on Galileo assessments continue to align well with the scores they will eventually achieve on the statewide assessment.

3. Forecasting Risk of Failing to Demonstrate Mastery on Statewide Test

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 assessments 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. Galileo® district-wide assessment results can provide an initial risk assessment based on a student's performance on a single district-wide test. As additional data become available for subsequent district-wide tests, risk assessments can be adjusted and refined. The Galileo *Risk Assessment Report* gives stakeholders real-time estimates of students' risk of not demonstrating mastery on statewide assessments so that intervention groups can be established and intervention efforts implemented. The following screen shot illustrates a sample *Risk Assessment Report*.

Galileo Risk Assessment Report – Confidential Screen Shot



Settings | Dashboard | Books | Help | Forum | Tech Support | Site Map | Logout

Risk Assessment Report

District: Mogollon Rim School District
Year: 2009 - 2010
School:
Class:

District: Mogollon Rim School District
(376 students have taken at least one of the tests listed below.)

Results From:
→ GCSD 3 Gr Math #1
→ GCSD 3 Gr Math #2
→ GCSD 3 Gr Math #3

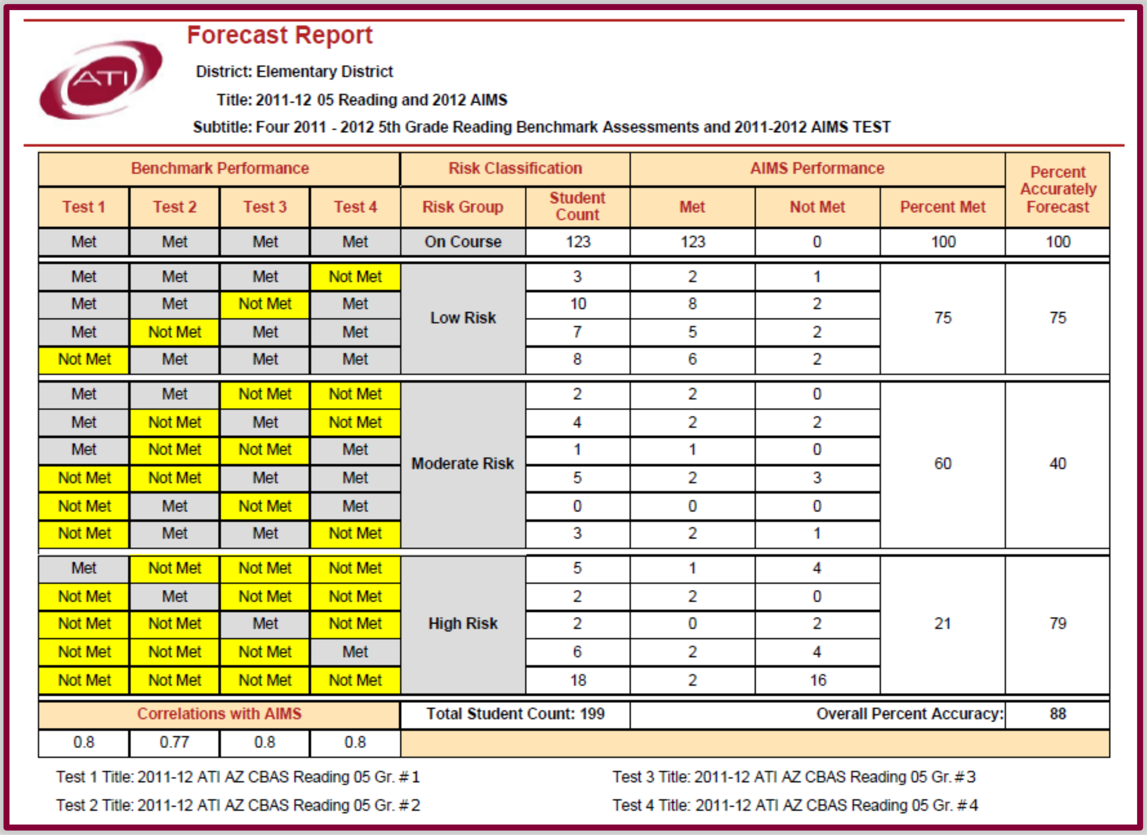
You can click on a school to view the detail for each class or click the count to view the detail for each class at the corresponding risk level. Use the browser's Back button to return to the previous level.

View Students	High Risk	Moderate Risk	Low Risk	On Course (minimal risk)
Brookfield Elementary School (49)	**19**	**3**	**15**	**12**
	38.78 %	6.12 %	30.61 %	24.49 %
New Oak Ridge School (84)	**14**	**13**	**18**	**39**
	16.67 %	15.48 %	21.43 %	46.43 %
Summerfield School (208)	**92**	**62**	**31**	**23**
	44.23 %	29.81 %	14.90 %	11.06 %
Williamsburgh Elementary School (39)	**12**	**12**	**11**	**4**
	30.77 %	30.77 %	28.21 %	10.26 %

Student risk-level classifications in the *Risk Assessment Report* are based on their performance on the set of district-wide assessments selected for inclusion when the report is run. For example, in the preceding screen shot, student risk-level classifications are based on performance on three third grade math benchmark assessments. Students who passed all three benchmark assessments are classified as “On Course.” Students who passed two of the three assessments are classified as “Low Risk.” Students who passed one of the three assessments are classified as “Moderate Risk.” Students who did not pass any of the three benchmark assessments are classified as “High Risk.” Further drill-downs from this view provide a list of students at each risk level as well as the list of the standards on which intervention efforts for each risk group should focus. Risk assessment information supports 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.

It is hoped that districts and charter schools rely on the *Risk Assessment Report* to plan and implement intervention efforts all year long. It is important, therefore, to monitor the accuracy of the forecasts made by the *Risk Assessment Report* on an annual basis. Like the correlation analysis discussed earlier, the forecasting accuracy analysis can only be conducted once the district/charter school has uploaded the student scores on the end-of-year statewide assessment. ATI conducts forecasting accuracy analyses for all districts/charter schools that elect to upload into Galileo® the results of the statewide assessment for individual students. The results of these investigations are presented as part of the Galileo *Forecast Report.* The *Forecast Report* illustrates the accuracy of Galileo risk forecasts for a district/charter school with respect to the observed student performance on the statewide assessment at the end of the year. Correlation coefficients indicating the relationship between student scores on each administered district-wide assessment and student scores on the statewide assessment are also presented in this report. A sample Galileo *Forecast Report* is illustrated in the following screen shot. This sample *Forecast Report* summarizes data from a representative district for four fifth grade reading benchmark assessments and the statewide fifth grade reading assessment (i.e., AIMS).

Sample Galileo Forecast Report – Confidential Screen Shot



Forecast Report

District: Elementary District

Title: 2011-12 05 Reading and 2012 AIMS

Subtitle: Four 2011 - 2012 5th Grade Reading Benchmark Assessments and 2011-2012 AIMS TEST

Benchmark Performance				Risk Classification		AIMS Performance			Percent Accurately Forecast
Test 1	Test 2	Test 3	Test 4	Risk Group	Student Count	Met	Not Met	Percent Met	
Met	Met	Met	Met	On Course	123	123	0	100	100
Met	Met	Met	Not Met	Low Risk	3	2	1	75	75
Met	Met	Not Met	Met		10	8	2		
Met	Not Met	Met	Met		7	5	2		
Not Met	Met	Met	Met		8	6	2		
Met	Met	Not Met	Not Met	Moderate Risk	2	2	0	60	40
Met	Not Met	Met	Not Met		4	2	2		
Met	Not Met	Not Met	Met		1	1	0		
Not Met	Not Met	Met	Met		5	2	3		
Not Met	Met	Not Met	Met		0	0	0		
Not Met	Met	Met	Not Met		3	2	1		
Met	Not Met	Not Met	Not Met	High Risk	5	1	4	21	79
Not Met	Met	Not Met	Not Met		2	2	0		
Not Met	Not Met	Met	Not Met		2	0	2		
Not Met	Not Met	Not Met	Met		6	2	4		
Not Met	Not Met	Not Met	Not Met		18	2	16		
Correlations with AIMS				Total Student Count: 199		Overall Percent Accuracy:			88
0.8	0.77	0.8	0.8						

Test 1 Title: 2011-12 ATI AZ CBAS Reading 05 Gr. #1

Test 2 Title: 2011-12 ATI AZ CBAS Reading 05 Gr. #2

Test 3 Title: 2011-12 ATI AZ CBAS Reading 05 Gr. #3

Test 4 Title: 2011-12 ATI AZ CBAS Reading 05 Gr. #4

As the *Forecast Report* illustrates, risk of failing to show mastery on the statewide assessment is related to student performance on the benchmark assessments. As the estimated risk for a student increases, the student is less likely to pass the statewide assessment. In the preceding sample *Forecast Report*, 100 percent of students classified as “On Course” and 75 percent of students classified as “Low Risk” passed the statewide assessment. On the contrary, only 60 percent of students classified as “Moderate Risk” and 21 percent of students classified as “High Risk” passed the statewide assessment. For purposes of evaluating the accuracy of Galileo® risk forecasts, students classified as either “On Course” or “Low Risk” are predicted to pass the statewide assessment while students classified as either “Moderate Risk” or “High Risk” are not predicted to pass the statewide assessment. In the preceding sample *Forecast Report*, the overall forecasting accuracy was 88 percent.

At the time of this writing, forecasting accuracy analyses have been conducted for the 2011-12 Galileo benchmark assessments administered by the first 33 districts/charter schools in Arizona, California, Colorado, and Massachusetts to provide ATI with their statewide test data for individual students. Figure 2 presents the results of these analyses. The results are based on 489 forecasting opportunities, where a forecasting opportunity is the student-level predictions made for the students within a given grade level in a specific district/charter school with regard to their likely performance on the statewide assessment in a given content area (i.e., one forecasting opportunity is for the third grade students in District A with regard to the statewide assessment in math). The results indicate that, as predicted, the majority of students who were classified as being “On Course” based on their performance on the Galileo benchmark assessments did in fact pass the statewide assessment, while the majority of those who had been classified as being at High Risk of not demonstrating mastery on the statewide assessment did in fact fail. The other two risk groups performed as expected as well. It should be noted that, if teachers and administrators are using the data provided by Galileo® benchmark assessments to implement effective interventions, many students who have been classified as being at some risk of failing the statewide assessment should pass it instead, thereby reducing the accuracy of risk assessment forecasts for the those student groups. We therefore consider a certain degree of inaccuracy in predictions of failure to be a sign of success.

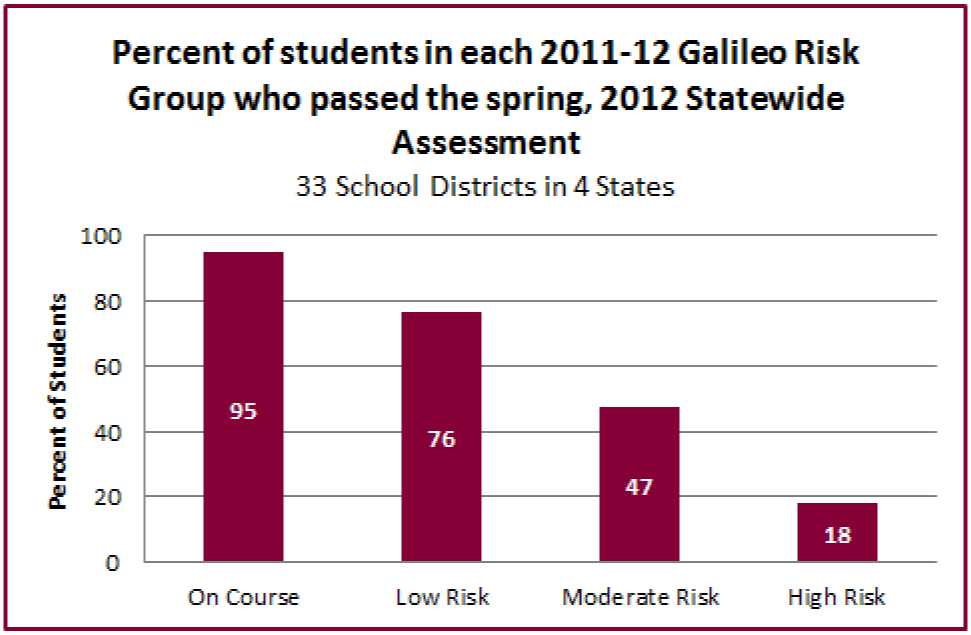


Figure 2
Forecasting accuracy of 2011-12 Galileo district-wide assessments with regard to student performance on the spring 2012 statewide assessment

C. A Look Under the Hood: ATI Item Development and Certification Procedures

ATI's *Assessment Planner* is used by districts and charter schools to provide ATI Educational Management Services staff with information to help ensure that benchmark assessments built meet district/charter school requests, particularly as they relate to inclusion of specific standards, test length, and test delivery dates. The following information describes ATI's item development and certification procedures and is intended to help the reader understand the procedures that are used to ensure the quality of the items comprising ATI benchmark assessments.

i. Guiding Standards for Developing Items

ATI's Assessment and Instructional Design Department constructs and certifies new items for ATI item banks on a continuous basis. ATI's approach to item construction includes strict conformance to detailed item specifications for *consistent measurement* of standards. Continuous item development increases the variety of items available for inclusion in district/charter school assessments. This is essential not only to reflect the broad range of district/charter school needs, but also because standards continually change as in the recent adoption of Common Core State Standards by many states. As standards change, curriculum also changes. To keep pace with this rapidly changing educational landscape, ATI continually updates not only the type of items and the content of its item banks, but also the psychometric analyses used to estimate item parameters including difficulty, discrimination, and guessing.

1. Certified Items

Only certified items can become part of a test generated to meet the specifications of a district/charter school. ATI's certification procedures promote high standards for item quality, minimize the likelihood of errors in test items, and provide increased variety in item selection. Item parameter estimates are saved to ATI item banks only for certified items following assessment with a significant number of students. The availability of item parameter estimates provides important information on item quality to be used in guiding test construction as described in Section II.B.

2. Suggestions for New Items

ATI welcomes suggestions for new items. Such suggestions help to ensure that assessments will meet district/charter school needs. ATI currently has one of the largest item banks in the nation containing approximately 118,000 items. Moreover, as mentioned previously, the banks are continuously growing. When making suggestions for new items, there are a number of considerations that the district/charter school may find useful. Suggestions for new items are likely to benefit the district/charter school most when they focus on the measurement of a specific skill or capability. If the district/charter school curriculum includes a particular capability that is not currently well represented in ATI item banks, suggesting the inclusion of additional items assessing the skill in question will be beneficial. Suggesting stylistic changes is generally less useful. ATI does develop items specifically aligned to the various styles of statewide assessments and currently is developing technology enhanced items to be used in the 2013-2014 academic year. It is important to ensure that items on local benchmark and formative assessments not be limited to a particular style. As items become closer in style to a statewide test, the danger that test outcomes will be contaminated by the phenomenon of "teaching to the test" increases. Benchmark and formative assessments are intended to measure the mastery of standards, not merely the ability to respond to items written in one particular style.

Finally, when making suggestions for new items, please keep in mind that in order to ensure adherence to guiding standards for item construction, review, and certification. ATI generally does not construct new items requested by individual districts/charter schools during the formal planning, construction, and review of benchmark assessments for that district/charter school. This policy benefits the district/charter school in a number of ways. It maximizes the quality of items included on an assessment because the included items will already have been certified and typically used successfully on many assessments. It increases the likelihood that the assessment will be effective in forecasting performance on statewide assessments because the included items will be drawn from a pool used effectively in previous forecasting initiatives, and it eliminates scheduling uncertainties because it ensures that the planning and construction process can proceed, following well established scheduling guidelines that have been proven to be effective for districts/charter schools.

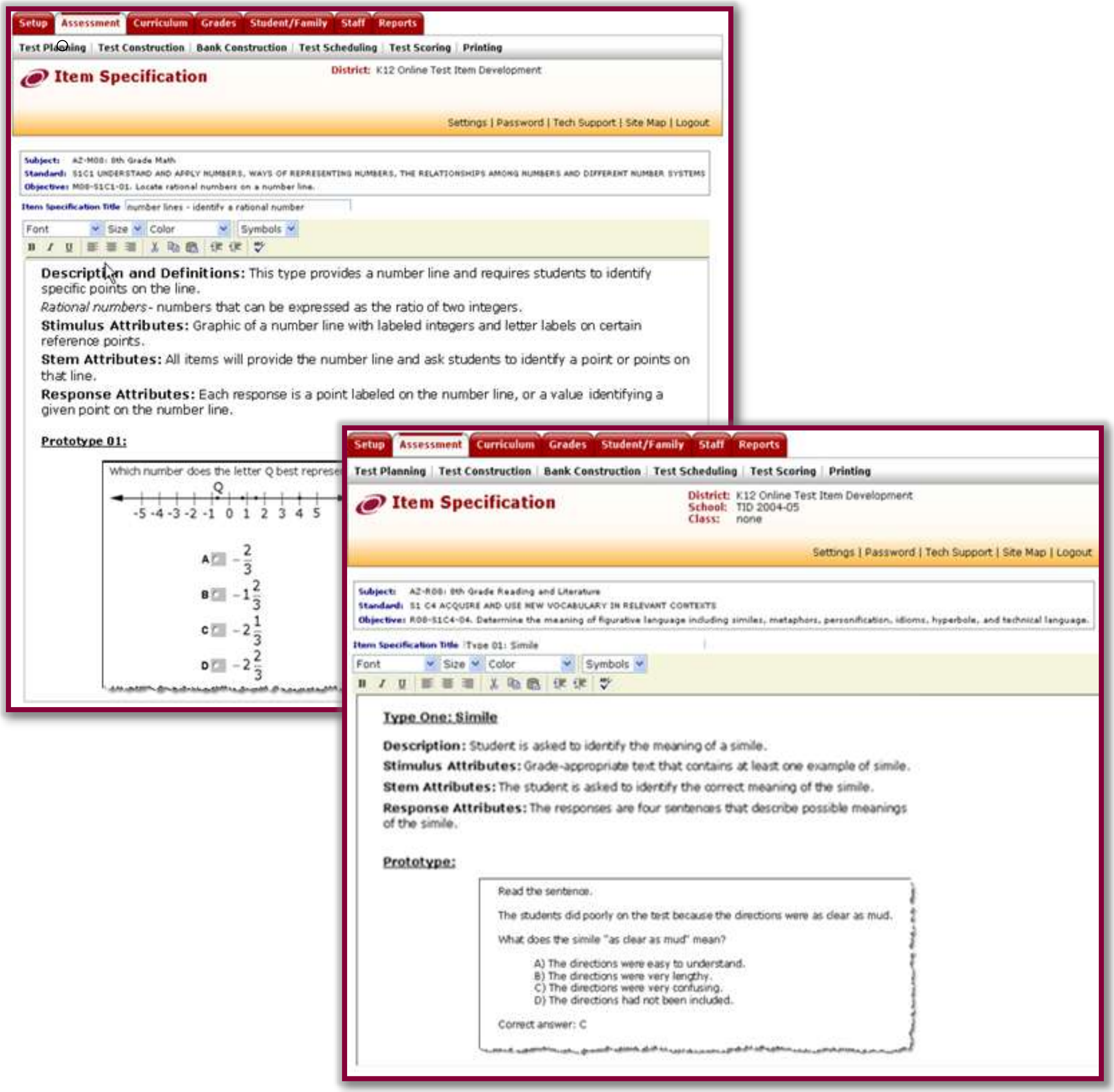
ii. Item Specifications

The item development process begins with the construction of ATI item specifications designed to guide item development for a particular standard. ATI item specifications include the following components:

- the standard that items conforming to the specification are designed to measure;
- the general description of the type of item to be covered by the specifications;
- the defining attributes of item components; and
- the sample item(s) demonstrating the specification.

Item specifications serve as a consistent point of reference to guide and evaluate the overall quality and utility of each individual item. To ensure that specifications are always accessible during item development, ATI includes them in the online item development environment used by ATI's Assessment and Instructional Design Department. This is illustrated by the screen shots that follow.

Item Specifications – ***Confidential Screen Shots***


Setup
Assessment
Curriculum
Grades
Student/Family
Staff
Reports
Test Planning | Test Construction | Bank Construction | Test Scheduling | Test Scoring | Printing
Item Specification
Settings | Password | Tech Support | Site Map | Logout
Font
Size
Color
Symbols

District: K12 Online Test Item Development

Subject: AZ-M08: 8th Grade Math
Standard: S1C1 UNDERSTAND AND APPLY NUMBERS, WAYS OF REPRESENTING NUMBERS, THE RELATIONSHIPS AMONG NUMBERS AND DIFFERENT NUMBER SYSTEMS
Objective: M08-S1C1-01. Locate rational numbers on a number line.

Item Specification Title number lines - identify a rational number

Description and Definitions: This type provides a number line and requires students to identify specific points on the line.

Rational numbers - numbers that can be expressed as the ratio of two integers.

Stimulus Attributes: Graphic of a number line with labeled integers and letter labels on certain reference points.

Stem Attributes: All items will provide the number line and ask students to identify a point or points on that line.

Response Attributes: Each response is a point labeled on the number line, or a value identifying a given point on the number line.

Prototype 01:

Which number does the letter Q best represe

A □ $-\frac{2}{3}$

B □ $-1\frac{2}{3}$

C □ $-2\frac{1}{3}$

D □ $-2\frac{2}{3}$

District: K12 Online Test Item Development
School: TID 2004-05
Class: none

Subject: AZ-R08: 8th Grade Reading and Literature
Standard: S1 C4 ACQUIRE AND USE NEW VOCABULARY IN RELEVANT CONTEXTS
Objective: R08-S1C4-04. Determine the meaning of figurative language including similes, metaphors, personification, idioms, hyperbole, and technical language.

Item Specification Title Type 01: Simile

Type One: Simile

Description: Student is asked to identify the meaning of a simile.

Stimulus Attributes: Grade-appropriate text that contains at least one example of simile.

Stem Attributes: The student is asked to identify the correct meaning of the simile.

Response Attributes: The responses are four sentences that describe possible meanings of the simile.

Prototype:

Read the sentence.

The students did poorly on the test because the directions were as clear as mud.

What does the simile "as clear as mud" mean?

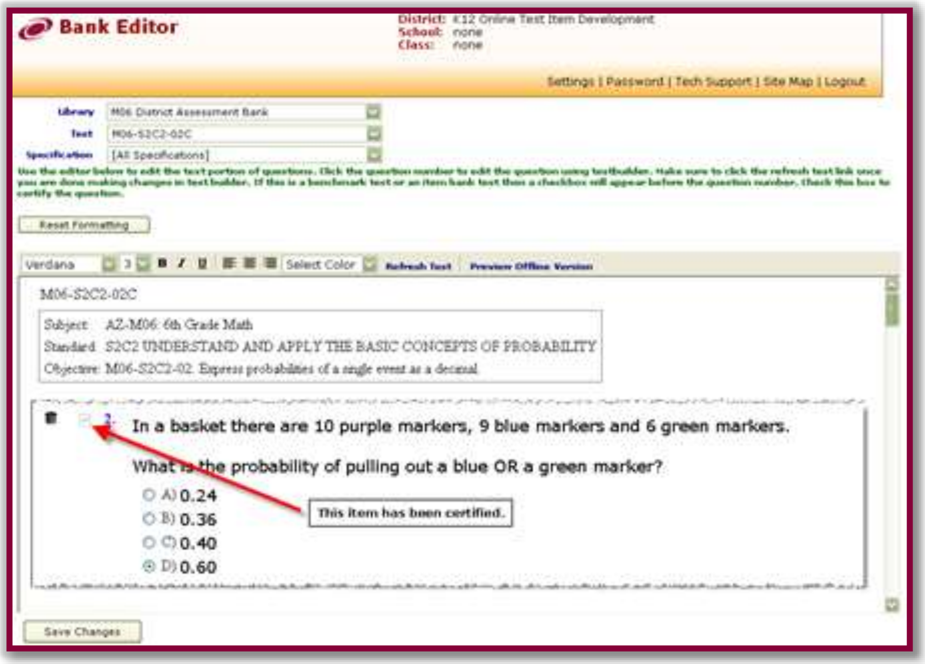
A) The directions were easy to understand.
B) The directions were very lengthy.
C) The directions were very confusing.
D) The directions had not been included.

Correct answer: C

iii. Item Review and Certification

Once an item is developed according to a specification, the item is subjected to ATI's internal screening process designed to identify any problems that might be present in the item. The screening process involves looking closely at all aspects of the item including the fit of the specification to the standard, the fit of the item to the specification, the clarity and appropriateness of the language in the item components, the fairness of the content, the characteristics of the response options, the graphics and fonts in the item and the correct answer indication. If the item fails on any of the screening components, the item is returned for specified revision. The process is repeated until the item is "accepted" and ready for certification, the final step in the item evaluation process. Certification involves a final item review in which alignment with standards and conformance of items to specifications guiding item development are again confirmed. The item certification process is completed online by education professionals in the Assessment and Instructional Design Department.

Bank Editor – Confidential Screen Shot



District: K12 Online Test Item Development
School: none
Class: none

Library: M06 District Assessment Bank
Test: M06-S2C2-02C
Specification: [All Specifications]

Use the editor below to edit the text portion of questions. Click the question number to edit the question using testbuilder. Make sure to click the refresh test link once you are done making changes in test builder. If this is a benchmark test or an item bank test then a checkbox will appear before the question number. Check this box to certify the question.

M06-S2C2-02C

Subject:	AZ-M06: 6th Grade Math
Standard:	S2C2 UNDERSTAND AND APPLY THE BASIC CONCEPTS OF PROBABILITY
Objective:	M06-S2C2-02. Express probabilities of a single event as a decimal

2. In a basket there are 10 purple markers, 9 blue markers and 6 green markers.

What is the probability of pulling out a blue OR a green marker?

○ A) 0.24
○ B) 0.36
○ C) 0.40
● D) 0.60

This item has been certified.

D. Putting It All Together: Steps in Benchmark Assessment Planning, Construction, Review, and Delivery

i. Step One: Initial Benchmark Preparation Communications with ATI

District/charter school leaders who will be responsible for oversight of benchmark planning, construction, and review will be identified by ATI Field Services Coordinators during initial implementation planning. These names will be provided to ATI's Educational Management Services staff who will be awaiting a call from the district/charter school assessment leader(s).

The leader(s) responsible for oversight of benchmark planning, construction, and review, can begin their part of the process by:

- becoming familiar with the contents of this resource document;
- familiarizing district/charter school staff involved in benchmark assessment development activities with the document's contents; and
- making phone contact with ATI Educational Management Services by calling 1.520.323.9033 or 1.800.367.4762.

During initial phone contact and in subsequent phone calls as needed, Educational Management Services staff will assist the district/charter in the completion of the *Assessment Planner*. The *Assessment Planner* must be completed in order for ATI to begin the process of constructing customized benchmark assessments.

ii. Step Two: Use of Assessment Planner to Define Assessment Goals

The online *Assessment Planner* makes it possible for a district/charter school to provide ATI with specific input related to the standards that are to be included in local benchmark assessments for each participating grade and content area. The planner helps to ensure that ATI constructed benchmark assessments are aligned with the selected standards and local benchmark goals. Specifically, districts/charter schools use the *Assessment Planner* to communicate to ATI:

- the requested delivery date for each assessment;
- the planned number of benchmark assessments;
- the specific standards to be measured on each benchmark;
- the desired number of items for measuring each standard on each assessment; and
- any comments regarding specific benchmark needs.

During the course of completing Step Two, ATI Educational Management Services staff will be available to assist with questions related to completing the *Assessment Planner* and to make recommendations as needed. As the *Assessment Planner* is completed, it is important to keep the following in mind:

1. Assessment Delivery Date

The timeline for benchmark development starts six weeks (i.e., 30 business days) prior to the time of the final delivery date indicated on the *Assessment Planner*. ***The final delivery date is the date the test is to be delivered to the district/charter school****; the final delivery* *date **is not** the date the district/charter school wants to administer the assessments.* When selecting a delivery date for assessments that will be administered offline, districts/charter schools need to be aware of how much time is needed to print the necessary number of test booklets and any vacation days that will be occurring during that time frame. For example, if a district is administering an offline assessment on October 15 and it will take two weeks to print the assessments, October 1 should be selected as the final delivery date if district printers will be available during that full time period. When selecting a delivery date for assessments that will be delivered only online, the delivery date needs to take into consideration the time the district/charter school will need to schedule the assessment.

2. Number of Assessments per Year

Districts/charter schools tend to prefer three to four benchmark assessments per year. Typically each of these assessments is designed to follow a period of instruction.

3. Standards Being Measured and Number of Items per Standard

An essential consideration in building reliable and informative benchmark assessments is the selection of the assessment content. Content selection is typically influenced by the district or charter school's approach to curriculum. Galileo® technology allows the district/charter school to align benchmark assessments to pacing guides or curriculum maps. Other curriculum-related factors influencing the choice of standards to measure include:

- If the approach to curriculum includes provisions for revisiting previously taught skills, the district/charter school may wish to repeat selected standards over two of more benchmark assessments. For example, a curriculum may call for a spiral approach in which previously taught skills are revisited multiple times.
- If the curriculum is designed to be sensitive to students reflecting a broad range of abilities, the district/charter school may wish to include skills that have not yet been fully targeted for instruction. For example, a curriculum may support repeated assessments of the same standards and coverage of standards that, while only being "introduced," are a topic of interest to the district/charter school for assessment purposes.

As mentioned previously, another essential consideration in building reliable and informative benchmark assessments is test length. Tests with at least 45 items are considered important toward the goal of achieving reliability. A second length consideration is keeping the test short enough to be manageable for the students taking the test. Districts/charter schools generally find that limiting reading and literature assessments to no more than 50 items and math assessments to no more than 60 items is helpful in managing administration of assessments.

An additional consideration in building benchmark assessments is ensuring that the test reflects the full range of abilities that students' evidence. Restrictions in the range of abilities assessed will reduce the magnitude of the relationship between the tests and other assessments including statewide tests. The consequences of range restrictions can be illustrated by considering an extreme case: If a test were so easy that every student got a perfect score, the correlation between that test and any other test would be zero.

The number of items selected per standard will need to vary depending upon the number of standards the district/charter school wishes to measure in a specific assessment. The larger the number of standards selected, the smaller the number of items per standard will need to be if the test is going to meet the test length guidance provided above.

iii. Step Three: Confirmation that the Assessment Planner is Complete and Obtaining Educational Management Services Feedback

Once the online assessment plan has been electronically submitted to ATI, Educational Management Services will send the district/charter school leader(s) confirmation via phone and/or email. The confirmation will generally be followed by Educational Management Services contact that provides feedback on the benchmark plans. The Educational Management Services review of the benchmark plans will be particularly attentive to:

- the length and content coverage in the planned assessments.
- the assessment characteristics, including the likelihood that the assessment can be administered within a reasonable amount of time.

Guidance from Educational Management Services staff with expertise in test development is provided to the district/charter school in order to help ensure that the benchmark assessments will have the desired reliability and validity as well as optimal utility for instructional planning, intervention, progress/outcome assessments and risk analysis related to student performance on statewide tests.

iv. Step Four: Development of a Benchmark Draft by ATI

- Seven weeks from the specified delivery date for final versions of benchmark assessments, Educational Management Services will confirm with the district/charter school leader(s) via phone and/or email that ATI is beginning the process of draft construction. At that time, Educational Management Services will also confirm that the district/charter school has finalized its benchmark plans and remind the district/charter school that at six weeks from delivery, no further changes in the *Assessment Planner* can be accepted without changing the test delivery date. That means that no additional standards can be included once the test construction begins at six weeks.
- Six weeks from the specified delivery date, ATI staff begins construction of the benchmark assessment.

v. Step Five: Guided Review of Benchmark Drafts

Approximately four weeks from the specified delivery date, ATI will provide the district/charter school with draft benchmark assessments for the specified content areas (e.g., math) and grade levels. Reviews of drafts are not required by ATI and are offered as an option for districts/charter schools wishing to do reviews. For those districts/charter schools choosing to review benchmark drafts, ATI provides specific guidelines for leader(s) to implement during this process.

It is suggested that requests for changes in draft assessments be approached cautiously. When considering an item replacement, it is important to ensure that the replacement item does not adversely affect the range of difficulty desired for the assessment and that it contributes to the measurement of the standards selected for assessment in the benchmark plan. The items available for benchmark assessment have been reviewed and certified and are aligned with standards reflecting expected levels of student mastery. When a replacement is contemplated, it is important to ensure that the replacement item will increase the quality of the benchmark assessment as a measure of valued educational goals that it is reasonable to expect students to obtain.

1. Guidelines for District/Charter School Review of Benchmark Drafts

- District/charter school leaders supervise the review process and are the primary contact with ATI Educational Management Services during the review.
- District/charter school leaders are responsible for the training of and guidance of reviewers.

  In this regard, please consider the following:

  Providing reviewers with knowledge concerning item development, item certification, and benchmark construction will help to establish consistent criteria for test review, encourage review objectivity, and enhance the value of the review for the district. ATI recommends that leader(s) request initial and final reviewers read and discuss this resource document prior to initiating the review process. Reviewers aware of the test construction, item construction, and validating procedures have a knowledge base from which to operate during the test review process.

  Requesting changes in draft assessments should be approached cautiously. When requesting an item replacement, it is important to ensure that the replacement does not adversely affect the range of difficulty desired for the assessment and that it does contribute to the measurement of the standards selected for assessment in the district/charter school benchmark plan.

- District/charter school reviewers use Galileo® K-12 Online test review tools to provide feedback to ATI Educational Management Services.

  During the review process, a district/charter school can accept the draft benchmark assessment as is or can request replacement of specific items (initial reviewers) or make item replacements (final reviewer only) in the draft.

  Final reviewer can make suggestions for item development in the comment box at the top of the *Final Review* interface within Galileo K-12 Online. Please note carefully the material in Section C of this document for detailed guidance concerning requests for item development.

- District/charter school personnel complete all steps within the test review process.

  Set the initial reviewers of the benchmark assessment. These individuals will then be able to login to Galileo, see an electronic version of the benchmark assessment, and review the test. The initial reviews are submitted to the final reviewer.

Set the final reviewer. Only one person can be assigned this task for a specific assessment. This individual will look over staff comments and submit a final test review to ATI.

Complete initial reviews. Those who were identified as initial reviewers will each complete a test review online.

Complete final review. The final reviewer responsibilities include reviewing the work of the initial reviewers and making the final determination of content to be accepted as presented on the assessment draft and/or completing any item replacements. If the final reviewer has suggestions for items that the district/charter school wishes to have added to the item banks for future assessments, these suggestions can be included in the comment box on the Final Reviewer interface.

Save the final review. ATI will then construct the final version of the test based on feedback from the final review.

vi. Step Six: Final Benchmark Assessment Delivered by ATI

Following the interchange of district/charter school input and ATI recommendations, ATI produces a final benchmark assessment for administration according to the district or charter school's benchmark final delivery schedule. In order to maintain the integrity of the decisions reached between the district/charter school and ATI during the draft review process, and to ensure the timely delivery of benchmark assessments, final benchmarks, unlike drafts, are not published for additional reviews and modifications.

III. Conclusion

ATI has created a unique process which enables cooperation between district/charter school partners and ATI Educational Management Services coordinators to build reliable and valid assessments including benchmark assessments and other types of district-wide assessments. Benchmark assessments that are aligned to local and state standards including Common Core State Standards can also be aligned to district/charter school mandated pacing guides, adding to the usefulness and validity of the assessments. These assessments can provide valuable data about student standards mastery, achievement, and growth throughout the year to be used by teachers, administrators, specialists, and parents in the context of initiatives to improve instructional effectiveness and student learning. Familiarizing district/charter school leaders and staff responsible for benchmark planning, construction, and review with the guidelines presented in this document is an important first step in ensuring the creation of valid, reliable customized benchmark assessments. ATI Educational Management Services coordinators will also provide valuable additional guidance and recommendations to districts/charter schools throughout the process.