Validity and Reliability of

Classworks Universal Screeners

Updated May 2018

Table of Contents

Executive Summary

Purpose

Classworks Universal Screeners are formal assessments used to measure readiness for grade level instruction, help identify baseline learning levels, and measure growth. The Universal Screeners were specifically designed for the purpose of screening students who may need additional intervention and can be used as part of the Response to Intervention (RtI) process.

In addition to reporting an overall scaled score based on the total test, Classworks provides student strengths and weakness for key strands. Key strands include a minimum of four test questions to provide a reasonable estimate of student strengths and weaknesses. This information, when used in conjunction with other data such as High Stakes Test results and classroom performance, can help provide a starting point for determining next steps.

Overview

Classworks Universal Screeners include multiple forms at each level for language arts and mathematics, grades K–10. The Universal Screeners are typically administered three times a year: at the beginning of the school year to assess readiness for instruction for all students, mid-year to measure progress for RtI tiers II and III, and end-of-year to measure overall growth for the year. Given that the test is primarily designed to identify readiness, the test includes multiple grade levels of content to allow sufficient reach for students who may be struggling.

The Universal Screeners are between 20 and 35 items in length depending on the grade level targeted, and must be administered in a single sitting. Two parallel forms of each Screener were developed; these forms measure similar content. The kindergarten level assessments are an exception to this approach, with two different forms reflecting earlier and later kindergarten content given the rapid development at the kindergarten level.

Overall test results are reported as a scaled score. Scoring on a vertical scale provides a single point of reference to compare individual student gains from one test administration to the next, within and across school years. Measuring growth vertically serves a dual purpose: to track learning gains for individual students and to determine whether learning must be accelerated.

Classworks Universal Screeners have been evaluated by the National Center for Response to Intervention (NCRTI), and they received the highest reliability ranking.

Universal Screener Quick Guide

Item	Description
Purpose	Measure grade level readiness, help identify baseline, measure growth
Grades	K–10 Math, K–10 Reading
Levels of coverage per test	Test includes multiple grade levels of content to allow sufficient reach to help identify strugglers (exception: Kindergarten)
Audio	Audio support available for all grades
Length of test	Must be taken in one sitting; 20–35 items depending on grade level/subject
Vertical scale?	Yes. All scores are vertically scaled from K–10 for longitudinal tracking.
Output from test	Average readiness scaled score of students by class, teacher, custom group, demographic, and/or grade level

Test Design

SEG Measurement (SEG) has been instrumental in the design, development, testing, and analysis of Classworks Universal Screeners. SEG is an assessment, measurement, and research firm that provides assessment design, development, and implementation services for K–12, higher education, and credentialing programs. They have delivered over 100 million assessments to tens of thousands of schools and colleges in all 50 states.

Classworks Universal Screeners were designed and built for the particular purpose they serve. For this reason, they meet all of the criteria that define quality screeners: the assessments are brief, reliable, valid, equated, and measured on a vertical scale.

SEG initially created the assessments by hand-selecting items for each level and form of the tests. Forms were then equated through field testing and calibration so that each measures the same sets of skills at the same level of difficulty. Individual test items and the assessments themselves were designed with diversity in mind: including populations of cultural and linguistically diverse students, and special needs students. Guiding principles for assessment design were integrated into the process, including ensuring all items are written in a clear, concise manner and free of age, gender, ethnic, religious, or disability bias.

There are two parallel forms for each test in grades K–10. For 2nd grade and above, the test questions include content from the target grade level as well as from two grade levels below the target. Given that the test is primarily designed to identify readiness, the test includes multiple grade levels of content to allow sufficient reach and enough content coverage for students who may be struggling. The tests include approximately 50% of the content from the target grade, approximately 25% of the content from the grade below, and approximately 25% of the content from two grades below. The 1st grade assessment contains content from both 1st grade and kindergarten. The kindergarten assessment contains content drawn only from kindergarten with two different forms reflecting earlier and later kindergarten content, given the rapid development at the kindergarten level.

Grade Level	Number of Test Questions Scored	Number of Test Forms	Source of Test Questions
K Early	15 Reading/Language Arts; 15 Mathematics	1	100 % early K content
K Late	15 Reading/Language Arts; 15 Mathematics	1	50% later K content; 50% early K content
Grade 1	20 Reading/Language Arts; 20 Mathematics	2	50% grade 1 content; 50% grade K content;
Grade 2	25 Reading/Language Arts; 25 Mathematics	2	50% grade 2 content; 25% grade 1 content; 25% grade K content
Grade 3	25 Reading/Language Arts; 25 Mathematics	2	50% grade 3 content; 25% grade 2 content; 25% grade 1 content
Grade 4	25 Reading/Language Arts; 25 Mathematics	2	50% grade 4 content; 25% grade 3 content; 25% grade 2 content
Grade 5	30 Reading/Language Arts; 30 Mathematics	2	50% grade 5 content; 25% grade 4 content; 25% grade 3 content
Grade 6	30 Reading/Language Arts; 30 Mathematics	2	50% grade 6 content; 25% grade 5 content; 25% grade 4 content
Grade 7	30 Reading/Language Arts; 30 Mathematics	2	50% grade 7 content; 25% grade 6 content; 25% grade 5 content
Grade 8	30 Reading/Language Arts; 30 Mathematics	2	50% grade 8 content; 25% grade 7 content; 25% grade 6 content
Grade 9	30 Reading/Language Arts; 30 Mathematics	2	50% grade 9 content; 25% grade 8 content; 25% grade 7 content
Grade 10	30 Reading/Language Arts; 30 Mathematics	2	50% grade 10 content; 25% grade 9 content; 25% grade 8 content

Vertical Scale and Item Bank Calibration

The vertical scale was developed through a linked testing design such that all items could be calibrated together and placed on the same continuum. The field test data was used to calibrate the items and tests. Calibration is a process that places all tests and all test items on a common scale. This was used to create a single common scale from grade K to grade 10. In this way, scores from the tests are comparable across forms of the test and over time. A given score will have the same meaning regardless of which form is administered and regardless of when the student takes the test.

The assessments developed include sets of overlapping items across test forms at the same level and across adjacent grade levels. This facilitates the calibration of the item bank. SEG calibrated the items using IRT (one parameter Rasch model) to create a common vertical scale across grade levels.

The raw number of correct answers reflects a particular Rasch score (ranging from -4 to +4), which is then translated to the final scaled score for reporting purposes. When the student completes his/her screener, the scaled score and key strand level performance feedback are immediately available for reporting. The approach taken in the calibration and scoring process provides Rasch extrapolated norms.

As a further measure to ensure that the test questions and assessments are technically sound and are performing as expected, SEG analyzes the data from the fall test takers each year.

Curriculum Advantage reviews the results from the fall to make sure the tests are performing well. SEG examines the statistics for the tests as a whole (e.g., average scores, distribution of scores) and the statistics for individual test items (e.g., question difficulty and the ability of the question to distinguish between different levels of student performance). Based on this analysis, Curriculum Advantage further refines the tests, revising and replacing questions as necessary.

During the 2014-2015 item analysis, Curriculum Advantage made the decision to update the Universal Screener. New items were created and field tested during the 2015-2016 school year and officially added to the assessment for the 2016-2016 school year. The Universal Screener updates overview, goal, and constraints can be found on addendum I.

Score Reporting

Score Reporting is designed to provide reliable information useful for understanding overall student readiness and estimated student strengths and weaknesses in specific strands measured by the test. Scores are based on scaled scores that allow all tests to be placed on a common scale regardless of which form is administered and at what grade level. Results are reported at the total test and key strand level. Strands assessed vary by grade level and subject of the assessment. This approach provides a reasonable balance between the need for information on student strengths and weaknesses the need for sufficient score reliability.

Raw scores are calculated as the total number of items answered correctly on the screener. Performance on the assessments is reported as a scaled score on a vertical scale ranging from 200 to 800 spanning across grades K–10. Feedback is also provided at the key strand level. (see Vertical Scale and Item Bank Calibration above).

These strands were determined based on an analysis of over 31 state standards and then re-examined with the introduction of the Common Core State Standards. Crosswalks are available to show the relationship between the Classworks strands and these state standards.

Reading:

- Grammar/Usage/Mechanics
- Reading Comprehension
- Study Skills
- Word Analysis
- Writing
- Writing Process

Math:

- Algebra
- Geometry
- Mathematical Processes
- Measurement
- Numeration
- Operations
- Patterns
- Statistics and Probability

Strands that are reported are required to include a minimum of four test questions to provide a reliable estimate of student strengths and weaknesses.

Curriculum Advantage establishes score ranges that reflect levels of student readiness on the assessments. There are various approaches that can be used to identify appropriate cut points defining levels of readiness. Below details the method SEG recommended for creating appropriate cut points.

Establishing Cut Scores

The cut scores for Classworks Universal Screeners for grades 3–8 were established using a two-stage standard setting process. In the first stage, a BookMarking Procedure (Cizek and Bunch, 2007) was applied. This was followed by a second stage, in which the stage one potential cut scores were reviewed in light of student performance data and expectations for student performance.

The BookMarking Procedure is an item mapping approach to standard setting developed in the 1990's (Cizek and Bunch, 2007). The BookMarking Procedure as employed for Classworks involves the review of an ordered test booklet containing all the items for a given test arranged in order of difficulty from easiest to hardest (Mitzel, H.C., Lewis, D.M., Patz, R.J., and Green, D.R., 2001). The difficulty values for this procedure were obtained from the Rasch item calibrations obtained from the original development of the screeners. Based on the procedures suggested by Mitzel, et al (2001), content experts reviewed the ordered item booklet and were asked to identify ("bookmark") the item representing the first item for which the minimally proficient student would be unlikely to answer the item correctly (less than 50% probability). The difficulty of the item identified served as the potential cut score emerging from stage one of the standard setting.

In the second stage, the potential cut scores produced in stage one of the process were reviewed against the distribution of scores from operational testing to evaluate the number and percentage of

students that would "pass" and the number and percentage of students that would "fail" the assessment based on the stage one potential cut scores. In some cases, the stage one potential cut score was raised or lowered based on the impact rates or expected performance for the students.

Item Development

The Classworks assessment item bank was developed by a team of content experts from a third-party developer, a leader in the creation of high-stakes content for assessments produced by states and testing companies. The test items have been reviewed and refined through a multi-step process involving members of this test development team.

The Universal Screeners are composed of 100% four-response-option multiple-choice type questions. The items were specifically developed for the Universal Screener or were selected and modified from the existing Curriculum Advantage item bank.

Guiding Principles of Item Construction

In order to ensure item reliability and validity, guiding principles were used in the item construction process.

Item Construction:

- Items are written in clear, concise language at the appropriate grade level
- Items are written without age, gender, ethnic, religious, or disability bias
- Each item set measures both basic knowledge and higher-order thinking skills
- Items adhere to the objectives being assessed
- Items are constructed in a consistent manner
- Item content is current and relevant to audience
- Items are written in the form of questions, avoiding open ended or negative stems

Item Response Measurement:

- Items show consistency of student response
- Results can be generalized to the population
- Items are calibrated to ensure that scores have similar meaning over time
- After calibration, items are placed on a developmental/vertical scale to allow for the accurate comparison of students over time and across use of the items
- Student performance can be predicted from item response
- Target goals and norms can be developed from item response measures

Questions/Stems:

- Stems and reading passages will be at grade-level readability and must assess the skill being tested according to the level of Bloom's indicated
- Stems are free of age, gender, ethnic, religious, or disability stereotypes or bias

- Stems are written in question format and do not require sentence completion, true/false, and fill-in-the-blank
- Each stem has only one correct answer

Answers/Distractors:

- Answers are presented in a multiple-choice format with four answer options
- Distractors are written in a logical order (alphabetical, chronological)
- Distractors are approximately the same length and must be grammatically parallel
- Distractors are plausible and should not contain grammatical clues
- Distractors address a variety of common errors rather than the same error
- Distractor rationale is provided for each answer choice

The test items are multiple-choice questions, offering an efficient and reliable way to assess students' knowledge and skills. All items have one single best answer and responses are scored as correct or incorrect. Multiple choice measures have advantages over other types of item response, in that they are capable of covering a large amount of content in a relatively short period of time. Moreover, they can achieve high levels of reliability, providing users with a consistent and stable measure of student knowledge and skills over time.

Test Validation

Following the creation of the tests, SEG conducted a second verification of the assessment items. The verification process consisted of a comprehensive alignment review to establish the validity of the assessment items and to determine if they were accurately aligned to the objectives they purport to measure.

Curriculum Advantage continues to partner with SEG to ensure that the tests themselves, as well as assessment-related decisions, are psychometrically sound. This ongoing process includes further statistical analysis, item calibration, adjustments to the cut scores on the vertical scale, and overall evaluation of the quality of Classworks Universal Screeners.

Field Testing and Analysis

To ensure that the test items and assessments are psychometrically sound, SEG analyzed the item and test performance data based on the field test to be conducted by Curriculum Advantage in the fall of 2009, the fall of 2010 and the fall of 2011. Curriculum Advantage collected information from approximately 200–300 students per test form the first year, with exponential increases in each of the following years. SEG analyzes the results each year, providing both test and item level analyses including:

- **Overall test and subtest statistics**
    - Mean
    - Standard Deviation
    - Reliability
    - SEM (Standard Error of measure)
    - Overall Model Fit
    - Frequency Distribution
- **Item statistics**
    - P Value (percent correct)
    - Point biserial correlation (measure of item discrimination)
    - Logit value from -3 to +3 (person and item independent measure of item difficulty)
    - Item Infit statistic
    - Item Outfit statistic

SEG reviews the item statistics, and any item that does not demonstrate suitable psychometric characteristics are recommended for replacement. These statistics help ensure on-going relevance and validity.

Here are some of the statistics SEG calculates:

Total Test Statistics

- **Average Score on the Assessment –** SEG computes the average (mean) score achieved by students taking the assessment. This helps us determine if the assessment is properly targeted to the level of the students assessed.
- **Variation and Distribution of Scores on the Assessment –** SEG calculates the amount of variability (standard deviation) in the test scores achieved by students taking the assessment. This is another indicator of how well the test is targeted to the level of students assessed.
- **Reliability –** SEG computes the reliability of the test to ensure that the test is consistently measuring the knowledge and skills measured by the assessment across forms of the test and is stable over time.
- **Score Accuracy –** Any assessment score is subject to variation when a student takes the test multiple times. SEG estimates the amount of variation expected for a student score (Standard Error of Measure; SEM); this is an indicator of score accuracy.

Individual Question Statistics

- **Question Difficulty –** SEG computes the percentage of students who answer the questions correctly; this is an indicator of the difficulty of the question
- **Question Differentiation –** SEG computes the relationship between student performance on each individual question and the assessment as a whole; this is an indicator of how well the question differentiates between those students who have the knowledge and skills measured by the assessment and those who do not have the knowledge and skills.

National Center for Response to Intervention Review

The National Center for Response to Intervention used data collected during the 2009–2010 and 2010–2011 school years to further evaluate the quality of Classworks Universal Screeners. Following the implementation of the final Universal Screener forms, performance on the screeners and high-stakes tests were used to investigate the validity and classification accuracy of the Universal Screeners.

Reliability

Test reliability refers to the test score consistency and accuracy. Reliability values range from 0 to 1.00, with higher values indicating higher reliability. Using the data collected from the multi-state field test, the average reliability for Universal Screeners for reading from grades K–10 was found to be 0.90. For mathematics in grades K–10, the average reliability coefficient is 0.88. These high internal consistency measures indicate that the Universal Screeners are able to provide a reliable measure of student performance in reading and mathematics.

Validity

Test validity refers to the appropriateness of the tests for its intended purpose. Evidence for validity of the tests is gathered from the item development and test development process as well as statistical analyses.

Classworks Universal Screeners were specifically designed for the purpose of screening students who may need additional intervention. The items and tests have been field tested and evaluated using Item Response Theory to ensure that the items and tests are performing as expected. The rigorous processes followed for item and test development provide support for the content validity of the Universal Screeners.

Performance on the Universal Screeners has been compared to other high-stakes tests to ensure that performance on the Universal Screeners is consistent with performance on other assessments. During the 2010–2011 school year, Classworks Universal Screener data and high-stakes test data from over 11,300 students in a large southern state were collected to evaluate the correlation between the Universal Screener scores and the high-stakes test scores.

Rules of Thumb – Armstrong (2006), reiterating the recommendations of Smith (1984) suggests the following rules of thumb for validity data examining one measure of a construct in relation to another measure of that construct:

- Over .50 excellent
- .40 to .49 good
- .30 to .39 acceptable
- Less than .30 poor

On average, the correlation between the Classworks Universal Screener scores and the high-stakes test scores was 0.46 for mathematics and 0.63 for reading. Further, the screeners were found to agree with other measures in classifying students as “not at-risk” 93% of the time in mathematics, and 97% of the time in reading. These correlations between two tests measuring similar constructs support the construct validity of the interpretation of the Universal Screener scores.

Classification Analyses

In addition to the reliability and validity of the measures, the Universal Screeners were also evaluated with regard to the accuracy of classifying students as at-risk in comparison to an independent measure. It is important that the screeners are able to appropriately identify students who are at-risk and those who are not at-risk. In particular, it is critical that at-risk students are properly identified as being at-risk to get the instructional help that they need.

In order to evaluate the classification accuracy, Classworks Universal Screeners classifications were compared to the classifications determined by performance on high-stakes state assessments in reading and math. The comparisons provided a classification of students into one of four cells in a "confusion matrix." Students could be classified as at-risk or not at-risk based on the passing status for each of the two assessments as Pass-Pass, Pass-Fail, Fail-Pass, or Fail-Fail. The classification analyses were performed by evaluating sensitivity and specificity.

Negative predictive power is a measure that estimates the accuracy of classifying students as "not at-risk." A useful screening tool should have very high negative predictive power such that at-risk students are not misidentified as not being at-risk. Using test data for more than 11,300 students, the Universal Screeners were found to have 93% and 97% negative predictive power for math and reading, respectively.

Classworks Universal Screeners Update

Technical Report

This document provides a summary of the tasks completed in updating the Classworks Universal Screeners for reading in math in grades K – High School.

August 2016

Contents

Overview

This document provides supporting information regarding the updates to the Classworks Universal Screeners in Reading and Math for grades K – 10 that will be in place officially for the 2016-2017 school year. This document provides the final plans that were executed between March 2015 and July 2016 and provides statistical information regarding the items and forms.

Goals and Constraints

The goals for this project were to modify the Universal Screeners to be more reflective of the latest multiple choice items and expectations of students in K-12 education, while at the same time keeping the Classworks Screeners consistent with the current forms. It was agreed that this project would include the development of 125 new Reading and 125 new Math items for use in the new Universal Screener forms. Further goals and guidelines are noted below.

- All items will be four-choice multiple choice with a single correct answer.
- There will be no audio or video passages associated with the items.
    - Grades K-2 forms will have text-to-speech support. Any new items written will get this applied so the entire field test form has this support.
    - There may be consideration for a passage to be a video or audio clip, if it can be fully owned and hosted (so as to avoid links expiring) and if the delivery can support it.
- The test lengths will remain consistent with the current test lengths of scoreable content (field test lengths will be longer).
- The majority of the items on the current screeners will remain on the new screeners.
- The new content should be as seamless as possible with the current content.
- As with the current forms, any strand with at least 4 items will be considered a key strand and will be linked to instructional content.
- All of the items must align to the current Classworks content hierarchy. (subject/grade/strand/skill/objective – listed in the Appendix)
    - There will be no changes to, combinations of, or additions to the strands, skills, or objectives.
- Field test forms will include the entire current form plus new items for field testing. Scoring during the field test time period will continue to be based on the scored items on the current forms.
- Reporting on student performance on the final new Screeners will need to be able to be comparable to historical performance on prior forms, whether by using the same scale or providing a translation of new to old scoring for comparison.
- The updates to the forms should be as seamless as possible.

Tasks

The following key tasks involved in updating the Universal Screeners are summarized in this report.

1. Investigating strands and expectations of other assessments

2. Finalizing plans for the updates to be made to the Universal Screeners

3. Selecting items for replacement

4. Developing new items

5. Producing new items

6. Creating field test forms and administering the field test

7. Analyzing field test data

8. Evaluating final test forms and scoring

Investigating strands and expectations of other assessments

As part of the initial planning stages, many assessments and standards were reviewed to gather information on the latest expectations of students in reading and math. This was to help meet the goal that the changes to the Universal Screeners would help to bring the forms more in line with expectations of other common assessments and standards.

The Common Core Reading/ELA strands are summarized in the following table.

Table 1: Common Core Reading/ELA Strands

Area	Strand	Grade										
		K	1	2	3	4	5	6	7	8	9	10
Reading	Literature - key ideas and details	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Literature - craft and structure	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Literature - integration of knowledge and ideas	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Literature - range of reading and level of text complexity	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Inf. Text - key ideas and details	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Inf. Text - craft and structure	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Inf. Text - integration of knowledge and ideas	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Inf. Text - range of reading and level of text complexity	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Foundational Skills - Print concepts	Y	Y	Y	Y	Y	Y					
	Foundational skills - phonological awareness	Y	Y	Y	Y	Y	Y					
	Foundational skills- phonics and word recognition	Y	Y	Y	Y	Y	Y					
	Foundational skills - fluency	Y	Y	Y	Y	Y	Y					
Language	Conventions of Standard English	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Knowledge of Language			Y	Y	Y	Y	Y	Y	Y	Y	Y
	Vocabulary Acquisition and Use	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Writing	Text types and purposes	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Production and Distribution of Writing	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Research to build and present knowledge	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Range of writing				Y	Y	Y	Y	Y	Y	Y	Y
Speaking and Listening	Comprehension and Collaboration	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Presentation of Knowledge and Ideas	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Literacy in History/Social Studies, Science, & Technical Subjects	Key ideas and details							Y	Y	Y	Y	Y
	Craft and structure							Y	Y	Y	Y	Y
	Integration of knowledge and ideas							Y	Y	Y	Y	Y
	Range of reading and level of text complexity							Y	Y	Y	Y	Y

The Georgia Milestone Assessment tests in grades 3 – 8 include the following high level skills for ELA:

- Reading and Vocabulary
- Writing and Language

The National Assessment of Educational Progress (NAEP) for Reading includes the following skills:

- Literary and Informational text
  - Locate and recall
  - Integrate and interpret
  - Critique and evaluate
  - Vocabulary

The Common Core Math strands are summarized in the following table.

Table 2: Common Core Mathematics Strands

Grade / Course	Strand	Grade													
		K	1	2	3	4	5	6	7	8	HS - Number & Quantity	HS - Algebra	HS - Functions	HS - Geometry	HS - Stats & Probability
K-8	Counting and Cardinality	Y													
	Operations and Algebraic Thinking	Y	Y	Y	Y	Y	Y								
	Number and Operations in Base 10	Y	Y	Y	Y	Y	Y								
	Number and Operations - Fractions				Y	Y	Y								
	Measurement and Data	Y	Y	Y	Y	Y	Y								
	Geometry	Y	Y	Y	Y	Y	Y	Y	Y	Y					
	Ratios and Proportions							Y	Y						
	The Number System							Y	Y	Y					
	Expressions and Equations							Y	Y	Y					

	Functions									Y					
	Statistics and Probability							Y	Y	Y					
HS Number and Quantity	The Real Number System										Y				
	Quantities										Y				
	Complex Number System										Y				
	Vector and Matrix Quantities										Y				
HS Algebra	Seeing Structure in Expressions											Y			
	Arithmetic with Polynomials and Rational Expressions											Y			
	Creating Equations											Y			
	Reasoning with Equations and Inequalities											Y			
HS Functions	Interpreting Functions												Y		
	Building Functions												Y		
	Linear, Quadratic, and Exponential Models												Y		
	Trigonometric Functions												Y		
HS Geometry	Congruence													Y	
	Similarity, Right Triangles, and Trig													Y	
	Circles													Y	
	Expressing Geometric Properties with Equations													Y	
	Geometric Measurement and Dimension													Y	
	Modeling with Geometry													Y	
HS Statistics and Probability	Interpreting Categorical and Quantitative Data														Y
	Making Inferences and Justifying Conclusions														Y
	Conditional Probability and Rules of Probability														Y
	Using Probability to Make Decisions														Y

The Georgia Milestone Assessment tests in grades 3 – 8 include the following strands for Math:

- Operations and Algebraic Thinking: Grades 3 – 5
- Number and Operations: Grade 3
- Number and Operations in Base 10 : Grades 4-5
- Number and Operations: Fractions: Grades 4-5
- Measurement and Data: Grades 3-5
- Geometry: Grades 3 – 8
- The Number System: Grades 6-7
- Ratios and Proportions: Grades 6 – 7
- Statistics and Probability: Grades 6 – 8
- Numbers, Expressions, and Equations: Grade 8
- Expressions and Equations: Grades 6 – 7
- Algebra and Functions: Grade 8

The National Assessment of Educational Progress (NAEP) mathematics assessment covers the following strands:

- Algebra
- Number properties and operations
- Measurement
- Geometry
- Data analysis, statistics and probability

Finalizing plans for the updates to be made to the Universal Screeners

We made recommendations for changes to the strands (particulars on consolidating, renaming, adding, or expanding) and after internal review of the impact on the system and benefits of making the changes, Curriculum Advantage determined that the strands will remain consistent between the current Universal Screener forms and the new Universal Screener forms. Rather than changing the strands, the focus is on increasing the quality of the items included within the strands.

Table 3: Classworks Reading/ELA Strands and Current Universal Screener Coverage

Grade	Grammar/Usage/Mechanics	Reading	Study Skills	Word Analysis	Writing	Writing Process	not covered - Listening/Speaking/Viewing	Grand Total
K		9	1	5				15
1	2	10		7		1		20
2	3	12	1	8		1		25
3	7	9	3	4	1	1		25
4	6	10	2	5	1	1		25
5	7	10	3	7	1	2		30
6	7	11	3	6	1	2		30
7	7	11	4	6		2		30
8	8	11	2	6		3		30
9	8	13	3	4		2		30
10	9	13	5	3				30

Table 4: Classworks Mathematics Strands and Current Universal Screener Coverage

Grade	Algebra	Concepts of Calculus	Geometry	Mathematical Processes	Measurement	Numeration	Operations	Patterns	Statistics and Probability	Trigonometry	Grand Total
K				2	3	5	2	2	1		15
1	1		5	4	4	4		1	1		20
2	2		4	3	5	4	2	2	3		25
3	2		2	2	6	1	5	2	5		25
4	1		5	1	3	4	4	1	6		25
5	5		8	4	4	1	2	1	5		30
6	6	1	8	3	1	3	3	1	4		30
7	6	2	6	2	4	2	2	1	5		30
8	8	1	8	1	4	1	2	1	4		30
9	8		7	5	2	1	1	1	4	1	30
10	8		6	5	4	1	1		4	1	30

The following decisions were made in conjunction with Curriculum Advantage with regards to the updates to the Universal Screeners:

- Each form would have approximately 20% of the form replaced with new items that align to the current Classworks objectives. (The objectives for each grade and subject were gathered through the Classworks item bank and included in Appendix A.)
- Items will be considered for replacement with a new item based on the quality of the current item, the importance of the objective measured, and the ability of the item to measure on-grade readiness.
- Items that are replaced may be replaced with a new item measuring the same objective, a different objective within the same strand, or an objective in a different strand that is in need of more coverage.
- In both reading and math, all new items will be single best answer multiple choice items.
- Items may be associated with one or more passages or images. Some items may need to be administered together in sequence as a set (i.e., a group of items that are all associated with the same passage(s)).
- Items will all be independent (not relate to or build on each other), even if they relate to the same passage or stimulus.
- New items may be used on multiple forms across or within grades (to follow similar overlap of current forms), but duplicate usage will only count as one item out of the 125 that will be developed.

- The field test forms will contain the entire current scoreable forms plus additional non-scored items for field testing. This will allow for the field test forms to continue to serve as live operational forms during the 2015-2016 school year.
- The new forms will maintain the current grade level coverage of the forms.
- All of the new items will be field tested to gather data.
- A linked form design with shared items will be used so that the entire pool of new items within a subject can be calibrated with the current pool.
- After the field test, the items and planned final forms will be evaluated.
- Scoring and comparability to the current forms will be evaluated to determine whether changes are warranted.

Table 5 shows the planned number of items developed for each grade (roughly 20% of each form). The actual item development matched these plans. Tables 6 and 7 show the breakdown of grade level coverage for each form, which remain consistent from the current scoreable items to the new scoreable items (after field testing).

Table 5: Number of New Items Per Form

Grade	**Reading**	**Math**
K	3	3
1	3	3
2	5	5
3	7	7
4	6	6
5	7	7
6	6	6
7	7	7
8	6 on one form, 7 on the other	6 on one form, 7 on the other
9	6	6
10	6	6
Total	125	125

Table 6: Item Grade Level Coverage on Reading Screeners

READING	Item Grade Level										
Form	K	1	2	3	4	5	6	7	8	HS	N
Grade K Reading Screener											
A	15										15
B	15										15
Grade 1 Reading Screener											
A	8	12									20
B	8	12									20
Grade 2 Reading Screener											
A	5	7	13								25

B	5	7	13								25
Grade 3 Reading Screener											
A		7	7	11							25
B		7	7	11							25
Grade 4 Reading Screener											
A			6	7	12						25
B			6	7	12						25
Grade 5 Reading Screener											
A				7	8	15					30
B				7	8	15					30
Grade 6 Reading Screener											
A					7	8	15				30
B					7	8	15				30
Grade 7 Reading Screener											
A						7	8	15			30
B						7	8	15			30
Grade 8 Reading Screener											
A							7	8	15		30
B							7	8	15		30
Grade 9 Reading Screener											
A								7	8	15	30
B								7	10	13	30
Grade 10 Reading Screener											
A								2	8	20	30
B								2	8	20	30

Table 7: Item Grade Level Coverage on Math Screeners

MATH	Item Grade level										
Form	K	1	2	3	4	5	6	7	8	HS	N
Grade K Math Screener											
A	15										15
B	15										15
Grade 1 Math Screener											
A	8	12									20
B	8	12									20
Grade 2 Math Screener											
A	5	7	13								25
B	5	7	13								25
Grade 3 Math Screener											
A		6	7	12							25
B		6	7	12							25

Grade 4 Math Screener											
A			6	7	12						25
B			6	7	12						25
Grade 5 Math Screener											
A				7	8	15					30
B				7	8	15					30
Grade 6 Math Screener											
A					7	8	15				30
B					7	8	15				30
Grade 7 Math Screener											
A						4	8	18			30
B						4	8	18			30
Grade 8 Math Screener											
A							7	8	15		30
B							7	8	15		30
Grade 9 Math Screener											
A								2	13	15	30
B								2	13	15	30
Grade 10 Math Screener											
A								2	8	20	30
B								2	8	20	30

Selecting items for replacement

Each current form was exported from the Classworks system into a separate Word document in preparation for review and update. For each form, the plans for numbers of items to be replaced and the blueprint for the form were noted in the document. Each current form was reviewed by experts to identify the specific items that would provide the most value by being removed from the form and replaced with a new item. The items were reviewed along multiple facets of quality including the general quality of the item, reflection of current expectations of the skill, importance and relevance of the item, and how well the item measures the objective within the skill/strand.

For each form, the items to be replaced were identified and item writing assignments were developed. In many cases, the new item would directly replace the current item with another item that better measured the objective within the strand. In some cases, it was determined that a different objective should be covered within the strand to better cover the focus of the particular strand.

Developing new items

After the items to be replaced were identified and the item needs were identified, item development began. Test development experts in math and reading developed the new items to meet the item specifications. New items were written to maintain the current style of the Universal Screeners while also representing newer ways of measuring the objectives.

The draft items were reviewed and edited for style, grammar, content accuracy, appropriateness, and perceived psychometric quality. The final 250 new items were then prepared for online production into the Classworks system.

Appendix B contains the alignment information for each of the new items.

Producing new items

Once approved internally, we individually entered the items into the Classworks item bank database. Each item was coded with its subject, grade level, strand, and skill as per the requirements of the system. The system generated a unique assessment system ID number for each item. The correct answer was identified and artwork was uploaded. The items were reviewed for proper rendering on the platform.

After the items passed through the internal production review, items were released to Curriculum Advantage for review by their content experts. In addition to the items being available online, details about the items were sent externally to assist in review and tracking. After review by Curriculum Advantage, the items were finalized by SEG and approved in the system by Curriculum Advantage content experts.

Once the items were approved in our local item bank in Classworks, Curriculum Advantage programmers worked to port the items into the official Classworks item bank and activate the items for use on the field test forms. During this process, the item IDs were slightly modified to ensure the item IDs were unique within the Classworks item bank while also allowing for tracking with the original IDs when the items were created. All of the new items in the official bank are in the 15,000s. For example, item ID 8 in the local bank is now 15008 and item ID 168 is now 15168.

Creating field test forms and administering the field test

In order to allow for continued production use of the Universal Screeners while also field testing the new items, the field test forms were developed to include the entire set of scoreable items on the current form as well as additional items for field testing that did not count towards the student's score. The field test items included items that would end up being on that official form as well as other linking items that would be dropped from the form. Items were placed strategically across forms so that all of the forms would be linked and that test data from students who were on grade, above grade, and below grade were exposed to the items. The following two tables summarize the plans for the field test forms. Appendix C contains the item level details on the field test forms.

Table 8: Field Test Plans for Reading/ELA

Form	Number of Scored Items on Current Form	Number of Non-Scored Items on Current Forms (these items will be dropped for new field test forms)	Current total number of items (scored and non-scored)	Number of Item Replacements (New Field Test Items that will eventually replace current scored items)	Number of linking items (additional non-scored linking items for field testing and calibration)	New Field Test Length (scored and non-scored items)	New Planned Screener Final Test Length (all scored items only, same number as current form scored item count)
Grade K Reading Screener							
A	15	5	20	3	3	21	15
B	15	5	20	3	3	21	15
Grade 1 Reading Screener							
A	20	5	25	3	3	26	20
B	20	5	25	4	2	26	20
Grade 2 Reading Screener							
A	25	5	30	5	3	33	25
B	25	5	30	5	3	33	25
Grade 3 Reading Screener							
A	25	5	30	7	3	35	25
B	25	5	30	7	3	35	25
Grade 4 Reading Screener							
A	25	5	30	6	4	35	25
B	25	5	30	6	4	35	25
Grade 5 Reading Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30
Grade 6 Reading Screener							
A	30	5	35	6	3	39	30
B	30	5	35	7	2	39	30
Grade 7 Reading Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30
Grade 8 Reading Screener							
A	30	5	35	6	3	39	30
B	30	5	35	7	2	39	30
Grade 9 Reading Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30
Grade 10 Reading Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30

Table 9: Field Test Plans for Math

Form	Number of Scored Items on Current Form	Number of Non-Scored Items on Current Forms (these items will be dropped for new field test forms)	Current total number of items (scored and non-scored)	Number of Item Replacements (New Field Test Items that will eventually replace current scored items)	Number of linking items (additional non-scored linking items for field testing and calibration)	New Field Test Length (scored and non-scored items)	New Planned Screener Final Test Length (all scored items only, same number as current form scored item count)
Grade K Math Screener							
A	15	5	20	3	3	21	15
B	15	5	20	3	3	21	15
Grade 1 Math Screener							
A	20	5	25	3	3	26	20
B	20	5	25	3	3	26	20
Grade 2 Math Screener							
A	25	5	30	5	3	33	25
B	25	5	30	5	3	33	25
Grade 3 Math Screener							
A	25	5	30	7	3	35	25
B	25	5	30	7	3	35	25
Grade 4 Math Screener							
A	25	5	30	6	4	35	25
B	25	5	30	6	4	35	25
Grade 5 Math Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30
Grade 6 Math Screener							
A	30	5	35	6	3	39	30
B	30	5	35	6	3	39	30
Grade 7 Math Screener							
A	30	5	35	7	2	39	30
B	30	5	35	7	2	39	30
Grade 8 Math Screener							
A	30	5	35	6	3	39	30
B	30	5	35	7	2	39	30
Grade 9 Math Screener							
A	30	5	35	6	3	39	30
B	30	5	35	7	2	39	30
Grade 10 Math Screener							
A	30	5	35	7	2	39	30
B	30	5	35	9	0*	39	30

*Grade 10 B form already has new field test items that are also on other forms/grades.

The field test forms were administered during the 2015-2016 school year as part of operational Classworks usage until sufficient data was collected for each form. Curriculum Advantage exported the field test data for analysis in June 2016.

Analyzing field test data

SEG prepared the field test data for analyses for multiple purposes: evaluating the item quality of the new items, evaluating the item quality of the current items that will remain on the forms, calibrating the new items into the current pools of active items, evaluating the difficulty of the test forms, and reviewing the vertical scaling across the forms.

The items were reviewed first in terms of percentage of students answering correctly. Any items that were answered by fewer than 25 percent correct were reviewed for accuracy. Items that have very few people answering correctly may simply be hard items, or they may be items that were miskeyed, did not render properly for answering correctly (particularly in the cases where images/graphs were required), or possibly had multiple correct answers. The point biserials were also reviewed for each item. The point biserial provides a measure of the relationship between performance on the item and performance on the form. All of the new items were determined to be functioning acceptably and no modifications or replacements were warranted. A small number of current items were flagged for content review internally at Curriculum Advantage for potential modification to improve the performance of the items. The items flagged for further review were items 8942, 13064, and 14628.

The detailed item statistics are provided in Appendix D.

The forms were reviewed to compare the overall difficulty of the planned new forms with the difficulty of the current forms. The new forms were found to be very consistent with the current forms as shown in Appendix E. These similarities were expected based on the final design and scope of the updates to the items on the forms.

Evaluating final test forms and scoring

After the field test data was evaluated and the definitions (item composition) of the new forms were confirmed, we evaluated the new forms to determine whether any changes to the scoring or use of the data would be warranted.

Using the data collected during the field testing, we calculated the estimated reliability of the new forms (including those items that will be scoreable on the final new forms). Reliability can be thought of as a measure of the consistency, stability, and accuracy of the scoring. Test scores with high reliability will produce similar scores for students if they were to retake the test without further instruction or time passing. Overall, the reliabilities for the new Universal Screeners are very strong. At the tails where there are fewer students taking the forms (specifically 10^{th} grade math), the reliabilities are a bit lower. The reliabilities are affected by the distribution of the scores and the students who took the test forms. It is expected that with additional test takers and a more consistent usage of the Screeners for those

forms, that we would see improved reliability for those forms where the reliability is currently a bit weaker than other forms.

Table 10: Form Reliability

	MATH	READING
KA	0.78	0.69
KB	0.88	0.86
1A	0.96	0.96
1B	0.77	0.82
2A	0.98	0.97
2B	0.87	0.97
3A	0.96	0.98
3B	0.97	0.97
4A	0.95	0.96
4B	0.91	0.89
5A	0.95	0.97
5B	0.93	0.93
6A	0.91	0.97
6B	0.92	0.94
7A	0.80	0.92
7B	0.89	0.93
8A	0.85	0.94
8B	0.85	0.93
9A	0.77	0.8
9B	0.51	0.75
10A	0.66	0.84
10B	0.48	0.82

The items were calibrated within subject across all grades and anchored to the current item pools. This was conducted in order to evaluate whether the items fit reasonably within the pool and whether changes to the vertical scaling were warranted. Given the consistency of the new forms with the current forms, it is recommended that the current scaling and reporting be continued. This will allow for longitudinal reporting in the system without changes to the system or increased complexity for teachers to interpret the results and make decisions. The item level logit and fit data from the vertical scaling is included with the item level statistics in Appendix D.

The new updated Universal Screener forms can be seamlessly put into production as planned and can continue to be used as an integral component of the complete Classworks system.