Design and Validity of

Summative Benchmark Assessments

Updated January 25, 2018

Table of Contents

Executive Summary

Purpose

Classworks Summative Benchmark assessments are intended to monitor student achievement of grade-level progress during the course of the school year and to assess outcomes at the conclusion of the year. The Summative Benchmark assessments are used primarily to measure growth over time and indicate whether the district, teachers, students, or sub-populations are meeting expected learning targets.

In addition to reporting an overall scaled score based on the whole test, Classworks displays 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.

Overview

Classworks Summative Benchmark assessments include multiple forms at K–8 for language arts and K–8 mathematics, as well as Geometry, Algebra 1, and English 1. The Summative Benchmark assessments are typically administered three times a year: at the beginning of the school year to determine a baseline of all students, mid-year to measure progress, and end-of-year to measure overall growth for the year. Given that the test is designed to measure grade-level mastery, the test focuses on content from only the grade level being measured.

The Summative Benchmark assessments are between 20 and 35 items in length, depending on the grade level targeted, and must be administered in a single sitting. Three parallel forms of each assessment were developed; these forms measure similar content.

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.

Benchmark Quick Guide

Item	Description
Purpose	Track student mastery of grade-level content during the school year, assess outcomes at the conclusion of the year
Grades	K–8 Math, K–8 Language Arts, Geometry, Algebra 1, English 1
Levels of coverage per test	Test includes on-grade level essential skills proven to be indicators of overall proficiency
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–8 for longitudinal tracking
Output from test	Average proficiency 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 Summative Benchmark assessments. 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 Summative Benchmark assessments were designed and built for the particular purpose they serve. For this reason, they meet all of the criteria that define quality summative assessments: brief, reliable, valid, equated, and measured on a vertical scale.

The Summative Benchmark assessments have been developed for grades K–8, Geometry, Algebra 1, and English 1. They are 20–35 items in length depending on the grade level targeted, to allow administration in a typical class period. There are three forms of each test, which allows teachers to assess students at multiple points during the school year. The Summative Benchmarks are developed to have similar content and to be statistically parallel to allow comparability of scores over time. This allows accurate measurement of growth from pre to post test.

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 by 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.

For all grades, the test questions are made up of on-grade level content only. The tests include on-grade essential skills proven to be indicators of overall proficiency.

Grade Level	Number of Test Questions Scored	Number of Test Forms	Source of Test Questions
Grade K	15 Reading/Language Arts; 15 Mathematics	3	On-grade level
Grade 1	20 Reading/Language Arts; 20 Mathematics	3	On-grade level
Grade 2	30 Reading/Language Arts; 30 Mathematics	3	On-grade level
Grade 3	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Grade 4	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Grade 5	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Grade 6	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Grade 7	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Grade 8	35 Reading/Language Arts; 35 Mathematics	3	On-grade level
Geometry	35 Geometry	3	Geometry
Algebra 1	35 Algebra 1	3	Algebra 1
English 1	35 English 1	3	English 1

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 through English 1. 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 Summative Benchmark, 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 also 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.

Score Reporting

Score Reporting is designed to provide reliable information useful for understanding overall student proficiency 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 by subject of the assessment. This approach provides a reasonable balance between the need for information on student strengths and weaknesses and the need for sufficient score reliability.

Raw scores are calculated as the total number of items answered correctly on the Benchmark. Performance on the assessments is reported as a scaled score on a vertical scale ranging from 1200 to 1800 spanning across grades K–8. Feedback is also provided at the key strand level (see Vertical Scale and Item Bank Calibration on page 6).

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

Reported strands 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 proficiency on the assessments. There are various approaches that can be used to identify appropriate cut points to define levels of proficiency. The following section details the method SEG recommended for creating appropriate cut points.

Establishing Cut Scores

The cut scores for Classworks Summative Benchmark assessments 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 Summative Benchmarks. 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 Summative Benchmark assessments are composed of 100% four-response-option multiple-choice type questions. The items were specifically developed for the Summative Benchmark 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/or 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 Summative Benchmark assessments.

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.

Independent Study: Correlation between Classworks Benchmark Assessments and High-Stakes Test Scores

Overview

Educators often want to predict student performance on state-wide high-stakes tests using data collected prior to the test; this information is used to target and monitor students who are at risk of not being successful on the state tests. To address this need, SEG Measurement (SEG) conducted research to evaluate models for predicting student performance on high-stakes tests using available Classworks data.

Methodology and Data Analysis

While the overall goal of the research was to identify a generalized model that could be used across tests and states, for purposes of the research, data from Georgia students was used and then the results were validated against data from Mississippi. The data included was 2010 and 2011 high-stakes test scores for students in grades 3 through 8.

The Classworks and high-stakes test data were merged into a single data file for use in this analysis such that each student record contained demographic data, Classworks instructional data, Classworks assessment data, and the high-stakes test score data for the student. Student records that did not include both Classworks and high-stakes test data were excluded from the analyses. In addition, Classworks test data outside of the 2010-2011 school year was excluded. Students who did not have Classworks instructional data were included as long as they had both Classworks assessment and high-stakes assessment data.

As a first step in the analysis, the overall relationship between the Classworks and Georgia CRCT scores was examined. The strength of the relationship between Classworks assessments and Georgia CRCT test scores was evaluated for Reading, English Language Arts, and Mathematics. Overall, the observed relationship between Classworks scores and Georgia CRCT scores was quite strong, suggesting that the development of a predictive model might be possible.

All of the tests included in this research are scored on different scales so to provide a common understanding and interpretation of different tests all test scores were converted to z-scores, a common scale often used for this purpose. Z-scores were calculated for the Classworks data using the large operational data set that included all users of Classworks. This provided a broader sampling, rather than limiting the scaling to one or two states. The z-scores for the CRCT scores were calculated using the students with Classworks data.

For each Classworks assessment subject and type, numerous regression analyses were run to investigate the independent variables that have a significant relationship with, or are significant predictors of, the Georgia CRCT scores for the respective subject. The use of linear regression allowed an investigation of the linear relationship that best fits the data and provides predictions regarding the dependent or criterion variable (in this case, high-stakes test score) given values for the independent variables or predictors (in this case, Classworks assessment data). Classworks Summative Benchmark scores and the prior CRCT score were consistently found to be statistically significant predictors.

Predictive models using the Classworks Summative Benchmark assessment data were created. For each grade, there is a model to predict the z-score on the CRCT in Reading, English Language Arts, and Mathematics using Classworks Summative Benchmark assessment score data.

In order to evaluate the utility of the models developed using the Georgia data for use with other high-stakes test data, the models were applied to Mississippi test data. For each subject and grade, both models were used to calculate predicted z-scores for Spring 2011. Specifically, for each student who had Classworks and 2010 MCT2 assessment data, the unique model specific to the particular subject and grade was used to calculate a predicted z-score on the 2011 MCT2 test. These predicted z-scores were then compared to the actual z-scores for Spring 2011 to evaluate the accuracy of the estimations. In addition, the predicted and actual z-scores were used to compare whether the predicted performance and actual performance agreed regarding the classification of each student as either proficient or not proficient.

Summary of Results

Using Classworks and Georgia high-stakes data, models were developed for predicting end-of-year Georgia CRCT scores. The predictive models consistently found that Classworks Summative Benchmark scores are statistically significant predictors of performance on the Georgia CRCT test.

Once predictive models had been identified, SEG converted the predictor tests to a common scale that could be generalized for use in predicting other state results based on the Georgia model. To validate the use of the models in other states, the models were applied in predicting similar test results in Mississippi. Those predicted scores were compared to the actual 2011 Spring MCT2 scores. On average, the predictions were accurate and the predictions of proficient status were accurate.

When administered with fidelity, Classworks Summative Benchmark assessment performance was found to be a significant predictor of performance on end-of-year high-stakes tests.