Evaluation of the Alignment Quality in the Georgia Milestones Assessment System in ELA, Mathematics, Science, and Social Studies

– Executive Summary –

February 2017

Ellen Forte, Ph.D.
Elizabeth Towles, Ph.D.
Elizabeth Greninger, Ph.D.
Erin Buchanan, M.A.
Lauren Deters, M.S.

edCount LLC®
because all students count

Table of Contents

List of Exhibits

Georgia Milestones Alignment Evaluation Report: Executive Summary

Evaluation Purpose

The Georgia Department of Education (GaDOE) commissioned edCount, LLC, to conduct an independent evaluation of the quality of alignment among its sets of academic standards and Georgia Milestones Assessment System to help ensure that the assessments yield meaningful, useful information for its stakeholders. The GaDOE intends to use the information gained via the evaluation to inform decisions about future item and assessment development and for federal peer review purposes. To support these purposes, the evaluation was designed to reflect *The Standards for Educational and Psychological Testing (The Standards; AERA/APA/NCME, 2014)* and to include analyses of six key aspects of alignment. The evaluation focused on both the design and development of the assessments and the outcomes of those processes in terms of the actual assessments as administered to students in Georgia.

This Executive Summary provides an overview of the methods used in this evaluation and the evaluation results.

Overview of GA Milestones Assessment System

The Georgia Milestones Assessment System is designed to measure how well students have acquired the skills and knowledge described in the Georgia state-mandated academic content standards. The assessments yield information on academic achievement at the student, class, school, system, and state levels. Scores on the Georgia Milestones EOC assessments are incorporated as final exams that count for 20% of course grades.

The Georgia Milestones Assessment System includes a series of criterion-referenced assessments in English language arts (ELA), mathematics, science, and social studies, organized in two major components. The EOG component refers to the assessments administered in grades 3-8 in each of these content areas, whereas the EOC component refers to ten course specific assessments administered to high school students. The EOC tests have been developed for the following courses: Ninth Grade Literature and Composition, American Literature and Composition, Analytic Geometry, Geometry, Coordinate Algebra, Algebra I, Biology, Physical Science, Economics, and United States History.

Key features of the assessments include the following:

- Integration of reading, language arts, and writing within a single assessment of English language arts
- Inclusion of constructed response items in the mathematics and English language arts assessments, in addition to the selected response items
- Inclusion of a writing component (in response to text) at every grade level and course within the ELA assessment
- Inclusion of norm-referenced items in every grade and content area to complement the criterion-referenced information and to provide a national comparison

- Transition to online administration over time, with online administration considered the primary mode of administration and paper-pencil back-up until transition is completed
- Eventual incorporation of technology enhanced items

Performance on the Georgia Milestones tests is reported on a scale of measurement specific to each grade, content area, and course. Performance on each Georgia Milestones test is also classified into one of four achievement levels: 'beginning learner', 'developing learner', 'proficient learner', and 'distinguished learner'. General test parameters for the Georgia Milestones Assessment System are displayed in Exhibit 1.

Exhibit 1. General Test Parameters for the Georgia Milestones Assessment System

	ELA	Mathematics	Science	Social Studies
Sections	3 sections, 1 of which includes an extended writing prompt	2 sections	2 sections	2 sections
Total number of items taken by each student				
	60	73	75	75
Criterion-referenced (CR)				
Total number of items	44	53	55	55
Total number of points	55	58	55	55
Breakdown by item type	40 Selected Response (1 point each; 10 of which are aligned Norm-referenced Test)	50 Selected Response (1 point each; 10 of which are aligned Norm-referenced Test)	55 Selected Response (worth 1 point each; approximately 10 of which are aligned Norm-referenced Test)	
	2 Constructed Response (2 points each)			
	1 Constructed Response (4 points)			
	1 Extended Response (7 points)			
Norm-referenced Test (NRT)				
Total number of items	20 (10 of which contribute to CR score)			
Embedded Field Test				
Total field test items	6	10	10	10

Methodology

edCount's approach to evaluating alignment quality within the Georgia Milestones Assessment System encompasses the collection and evaluation of a comprehensive body of evidence that itself aligns with the demands of both the federal peer review criteria for alignment and, even more importantly, *The Standards* (AERA/APA/NCME, 2014). Background on this approach is provided in the white paper on alignment commissioned by the Council of Chief State School Officers (Forte, 2016).

The evaluation framework encompasses 12 evaluation questions, two for each of six connections in the path from academic content standards to assessment scores. These connections and associated evaluation questions are:

1. The domain definitions (measurement targets) to the academic content standards.
    a. How were the measurement targets established to reflect the full depth and breadth of the academic content standards? Was this a reasonable and sound process?
    b. How well do the measurement targets address the full depth and breadth of the academic content standards?

2. The item specifications to the domain definitions.
    a. How were the task models and item templates developed to reflect the measurement targets? Was this a reasonable and sound process?
    b. How well do the task models and item templates reflect the measurement targets?

3. The assessment blueprints to the domain definitions.
    a. How were the blueprints developed to reflect the measurement targets? Was this a reasonable and sound process?
    b. How well do the blueprints reflect the measurement targets?

4. The ALDs to the blueprints.
    a. How were the ALDs developed to reflect the measurement targets? Was this system reasonable and sound?
    b. How well do the ALDs reflect the measurement targets?

5. The items to the blueprints and item specifications.
    a. How were the items developed to reflect the measurement targets? Was this system reasonable and sound?
    b. How well do the items reflect the measurement targets?

6. The sets of items contributing to a student's score to the blueprints.
    a. How were the forms and scoring rules developed to reflect the measurement targets? Was this system reasonable and sound?
    b. How well do the sets of items that contribute to students' scores reflect the assessment claims and the measurement targets?

This evaluation included six studies, one to address each of the six connections noted above. For each of the six studies, edCount's content and measurement experts leading the alignment study teams for ELA, mathematics, science, and social studies first conducted a thorough review and evaluation of the Georgia Milestones Assessment System design and development process to address the "a" questions, above. The purpose of this component of the evaluation was to determine if the assessment design and development process seems reasonable, adheres to standards of best practice, and is likely to yield assessments that provide scores that can be interpreted as intended (Forte, 2016).

To address the "b" questions, above, edCount's onsite alignment study team leaders trained and facilitated groups of expert panelists to evaluate the Georgia Milestones Assessment System outcomes. This evaluation serves the important purpose of confirming that the products resulting from the assessment design and development process are in alignment with the foundational logic and purpose for which the assessment system is built.

In study 1, edCount researchers examined how the GaDOE designed and developed the domain definitions from the standards and evaluated the degree to which the domain definitions address the full breadth and depth of the academic content standards. In study 2, edCount researchers reviewed how the GaDOE created item specifications to guide development of items that reflect the standards and rated the degree to which the item specifications represent the academic content standards and Depth of Knowledge (DOK) indicated in the standards documents.

To evaluate the blueprints in study 3, edCount researchers examined the documentation describing how the GaDOE developed blueprints to reflect the domains and then examined the relationship between the domain definitions and the blueprints as reflected by independent expert raters. Researchers considered how well the blueprints represented the standards in terms of content and skill match and in cognitive complexity.

Study 4 focused on the achievement level descriptors and the degree to which they reflect the standards via the domain definitions. edCount researchers examined the documentation that describes how the GaDOE designed and developed the ALDs as well as the ALDs themselves. In study 5, edCount researchers examined how the GaDOE developed the test items and rated actual items from an operational form. Independent panelists associated each item with the academic content standard and element(s) of the academic content standard to which the item aligned and rated the cognitive complexity using a common rubric. In study 6, researchers reviewed scoring rules and interpretive guides and evaluated the sets of items that contributed to students' test scores.

Results

Results for each of the studies are summarized in Exhibit 2; results are organized to correspond to each of the four traditional aspects of alignment (Categorical Concurrence, Range of Knowledge, Balance of Representation, and Depth of Knowledge) in Exhibit 3 to facilitate readers' interpretation of evaluation results in those terms.

The results of the six studies indicate that the GaDOE has engaged in a test and item development process that meets professional standards for quality and rigor and that the EOG and EOC assessments in its Georgia Milestones Assessment System adequately reflect the Georgia state-mandated academic content standards. The GaDOE is to be commended on voluntarily embarking on an extensive, comprehensive evaluation process.

Exhibit 2. Summary of Results by Study

Questions	Results based on evaluation of the process	Results based on evaluation of the outcomes
Study 1: Relationship between measurement targets and academic content standards		
1. How were measurement targets established to reflect the full depth and breadth of the academic content standards? Was this system reasonable and sound? 2. How well do the measurement targets address the full depth and breadth of the academic content standards? Note: GaDOE uses a traditional assessment development approach and has developed domain definitions to identify the content and skills that are eligible for assessment. GaDOE does not use the terms "claims" and "measurement targets", which are associated with evidence-centered design practices, but the domain definitions are considered to be the measurement targets for evaluation purposes.	The GaDOE developed domains definitions to reflect the academic content standards in all four content areas with the intentional exception of the Speaking and Listening strands in ELA.	ELA: The domain definitions include four of the six strands in the ELA standards. Two strands, speaking and listening, are excluded from eligibility for assessment because the GaDOE and its stakeholders determined that these are best assessed at the classroom level. Researchers suggest that the GaDOE consider providing guidance to support these classroom-based assessments. Results for the other content areas indicate that the full range of domains are included in the domain definitions.

Questions	Results based on evaluation of the process	Results based on evaluation of the outcomes
Study 2: Relationship between measurement targets and item specifications		
1. How were the item specifications developed to reflect the measurement targets? Was this system reasonable and sound? 2. How well do the item specifications reflect the measurement targets?	The GaDOE engaged in a comprehensive process involving multiple stakeholders to develop its test and item specifications in all content areas.	The clarification of task, item stimuli, and response attributes within the item specifications define and describe the content of the test, the item formats, and item properties.
Study 3: Relationship between measurement targets and blueprints		
1. How were the blueprints developed to reflect the measurement targets? Was this system reasonable and sound? 2. How well do the blueprints reflect the measurement targets?	The GaDOE developed its blueprints to reflect its domain definitions and involved external experts in this process.	The blueprints represent adequate score points in each domain (Categorical Concurrence) in each content area. In Social Studies, the large number of standards in many domains make it impossible to reflect the range of standards when evaluated by mere number and percent. The GaDOE is encouraged to establish measurement targets to reflect the range of concepts within these standards that are eligible for measurement. The test blueprints in all content areas sampled a variety of academic content standards addressed during instruction. The average DOK of the standards as rated by panelists corresponds to the range targeted in the blueprints.

Questions	Results based on evaluation of the process	Results based on evaluation of the outcomes
Study 4: Relationship between measurement targets and achievement level descriptors		
1. How were the ALDs developed to reflect the measurement targets? Was this system reasonable and sound? 2. How well do the ALDs reflect the measurement targets?	The GaDOE developed the ALDs with the participation of its stakeholders to reflect the expectations inherent in its standards. The GaDOE commissioned a third party review of the ALDs and used the results of this review to refine its ALDs. The GaDOE conducted standard setting with a large participant group using well-established methods. The GaDOE has conducted or commissioned several reviews of its processes and outcomes during the development of its assessments and uses this information to adjust its practices, where necessary. The GaDOE may wish to examine how the items on each assessment reflect the range and progression of knowledge and skills expressed in its ALDs.	Across grades and content areas, the ALDs reflect the content of the domains. Across grades and content areas, items reflect the ALDs such that the ALDs describe performance that students can demonstrate on the assessments. The GaDOE may wish to review the items not mapped to ALDs. The GaDOE may wish to review item writing specifications and blueprints to ensure that each form includes items that represent all four achievement levels in a pattern that best suits intended score interpretations.

Questions	Results based on evaluation of the process	Results based on evaluation of the outcomes
Study 5: Relationship among assessment claims/measurement targets and items		
1. How were items developed to reflect the measurement targets? Was this system reasonable and sound? 2. How well do the sets of items that contribute to students' criterion-referenced scores reflect the targets?	The GaDOE has establish strong technical documentation of the processes it uses passage and item development, including the training and guidance provided to item writers as well as their qualifications including the specific topics addressed during the passage and item writing trainings with regard to bias and sensitivity issues and adherence to the principles of universal design for assessment. The GaDOE regularly conducts content and bias review meetings with external experts to ensure that the test items used to contribute to students' CRT scores are aligned to the academic content standards. The GaDOE includes elements of the universal design for assessment through the test development process, most notably in terms of accessibility.	When rated independently, the items that appear on the EOG and EOC assessments reflect a high rate of match to the standards and DOK values of record, which indicates a strong connection between the items of the Georgia assessments and Georgia's academic content standards.
6: Relationship among assessment claims/measurement targets and the items that contribute to students' CRT scores		
1. How were the forms and scoring rules developed to reflect the measurement targets? Was this system reasonable and sound? 2. How well do the sets of items that contribute to students' scores reflect the assessment claims and measurement targets?	The GaDOE engages in and provides sufficient documentation of sound protocols for monitoring and quality control procedures in regard to scoring, processes for the development of scales for the purposes of score reporting, and processes for classifying students into ALD levels.	The operational forms, when evaluated independently, include sufficient score points for domains. In ELA, mathematics, and science, the operational forms reflect most standards within each domain and all domains are represented in proportion to the expectations in the blueprints. Cognitive complexity ratings indicate that the operational forms corresponded with the DOKs of record in the majority of strands across content area and grades/tests.

Exhibit 3. Summary of Results by Traditional Alignment Aspects

Aspect	Study 3: How well Blueprints reflect the Standards	Study 6: How well Operational Forms reflect the Blueprints	Results
Categorical concurrence *Is each reporting category (domain) associated with at least six score points?*	Count number of score points as indicated in blueprint for each domain	Count number of score points as indicated by panelists' ratings of item match to standards, aggregated to the domain level	With only one exception (Geometry in grade 3 Mathematics), all domains were associated with at least six points in both the blueprint and on the operational form.
Range of Knowledge *Does each domain represent at least 50% of the standards it encompasses?*	a) Calculate numbers of standards within strands and domains meant to be associated with items as indicated in the blueprint. b) Calculate the proportion of the standards within strands and domains these represent.	b) Calculate the proportion of the standards within strands and domains as represented by the blueprint (GA blueprints include all standards). c) Calculate numbers of standards within strands and domains associated with items as indicated by panelists' ratings of item-to-standard match.	Could not evaluate the blueprints against the standards because the blueprints include all standards as eligible for inclusion on the assessments. In ELA, Mathematics, and Science, items reflect at least 50% of the standards in most domains. Due to a large number of standards in some grades and courses in Social Studies, fewer than 50% of these standards were associated with items on the assessments.
Balance of Representation *Does the proportional representation of the domains reflect the standards and the blueprints?*	a) Calculate the proportion of all standards that the standards within each domain represent. b) Calculate the proportion of all standards that are associated with items in the blueprint and weight by score points.	b) Calculate the proportion of all standards that are associated with items in the blueprint and weight by score points. c) Calculate the actual distribution of score points based on panelists' ratings of the items-to-standards matches.	The proportional representation of domains as determined by panelists' ratings of individual items matched the proportional representation of domains indicated in the blueprints for all content areas. There were no instances in which the blueprint was either underrepresented or overrepresented by score points on an assessment.

Aspect	Study 3: How well Blueprints reflect the Standards	Study 6: How well Operational Forms reflect the Blueprints	Results
Depth of Knowledge	a) Calculate the average DOK of the standards at the strand and domain levels b) Calculate the average DOK of items at the strand and domain levels as indicated in the blueprint Compare a and b Could not do this for GA so just indicated the average DOK at the strand and domain levels	b) Calculate the average DOK of items at the strand and domain levels as indicated in the blueprint c) Calculate the average DOK of items at the strand and domain levels as indicated by panelists' ratings of item DOKs Compare b and c For GA had to use standard DOK ratings rather than the blueprint for comparison.	The DOK of items as rated by panelists was generally lower than the DOKs of record, but typically not significantly so. In ELA, panelists' DOK ratings were more than .5 lower than the DOKs of record at the strand level in only 9 of the 45 strands. In Mathematics, panelists' ratings of DOK were more than .5 lower than the DOK of record for 17 of the 76 strands and higher than the DOK of record in one strand. Panelists' ratings of DOK strongly matched the DOKs of record in Science and Social Studies.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Forte, E. (2016). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Technical Issues in Large Scale Assessment SCASS of CCSSO.