Validity and Reliability for the 2015-2016 Georgia Milestones Assessment System

The Georgia Department of Education (GaDOE) oversees the development of the Georgia Milestones Assessment System and adheres to the *Standards for Educational and Psychological Testing* (2014) as established by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The intent of these standards is "to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (AERA, APA, NCME, 1). Key to any assessment is the adherence to the *Standards* that address the issues of validity and reliability. While validity is the most important consideration in the test development process, a test cannot be valid without a high degree of reliability.

Validity

According to the *Standards*, "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (11). It is important to understand some key elements of validity. First, validity exists in context. A test may have a high degree of validity for one proposed use, but less validity for another. Second, validity is a matter of degree and is not an "all or nothing" condition. Finally, validity is associated with a multi-faceted process and collection of evidence over time. Questions of validity cannot be accurately summed up in a single statistic. Ultimately, the answer to a validity question for a test rests in careful documentation of the test development process. The *Standards* include 25 separate standards related to validity; however, these standards are not prescriptive for each test but should be considered as to whether they are applicable for the particular test under consideration. This brief will succinctly describe the major evidences of validity for the Georgia Milestones Assessment System.

One of the first pieces of evidence for establishing a test's validity is a clear identification of the purpose of the test. In the case of the Georgia Milestones assessments, the state legislature has identified the purpose to be a measure of how well students have mastered the state's content standards (O.C.G.A. § 20-2-281). The Georgia Milestones assessments are mandated by state law and are designed to measure how well students acquire the skills and knowledge described in the state's mandated rigorous content standards in English language arts, mathematics, science and social studies in grades three through eight and in selected high school courses. In addition to measuring how well students acquire the skills and knowledge described in the state's content standards, the Georgia Milestones assessments have the additional goals of identifying the areas where the students need improvement, informing various stakeholders of the progress toward meeting academic achievement standards of the state, meeting the requirements of the federal accountability, and gauging the overall quality of education in the state of Georgia. The assessments yield information on academic achievement at the student, class, school, system, and state levels. The evidence then for the validity of Georgia Milestones relies primarily on how well the assessment instrument matches the intended content standards and how the score reports inform the various stakeholders – students, parents, and educators – about the students' performance.

Therefore, the test development cycle for the Georgia Milestones must begin with the state's mandated content standards. Because the GaDOE believes that the content standards reside both in the approved published documents as well as the way they are enacted in the classroom, the test development process also relies heavily on the inclusion of educators from around the state. Once the purpose of the test is established, committees of educators are formed to review the content standards and establish which concepts, knowledge, and skills will be assessed and how they will be assessed. The results of this review produce several key documents. The *test specifications* indicate which standards can and will be measured and how they will be represented on the assessment. In conjunction with the test specification, the *content domain specifications,* along with the *test blueprint,* indicate how specific standards or elements will be grouped into reporting categories.

From these documents, an additional document is developed that provides more details for the item writing phase. Moreover, *test item specifications* are produced which give additional detail about what kinds of items will be written. This document typically identifies the item format, content scope and limits, and cognitive complexity. For example, item specifications for an English language arts assessment may address the genre, complexity, and/or length of literary passages to be produced. All of these activities are conducted by the Department with the assessment contractor with substantial involvement by curricular specialists and Georgia educators. The content domain specifications are then converted into a publically posted document known as the Georgia Milestones Assessment Guides so that all stakeholders are informed of the test's content and method of assessment. This document organizes the assessed content under the test domains, the structure in which the test results will be reported. In addition, Georgia Milestones Test Blueprints and Content Weight documents are posted on GaDOE's website to show the relative proportion of items by domain that are included on each content area test. These documents and the inclusion of Georgia educators serve as one piece of evidence of the Georgia Milestones validity as a measure of the state's content standards. To find the assessment guides, blueprints, and content weights for the End-of-Grade Assessments, visit http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOG-Resources.aspx and for the End-of-Course Assessments, visit http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOC-Resources.aspx.

Once this is accomplished, items are written by qualified, professional assessment specialists specifically for Georgia tests. Committees of Georgia educators review the items for alignment with the curriculum, suitability, and potential bias or sensitivity issues. The review committees have the authority to accept the item as is, revise it, or reject the item. Items that are accepted are placed on field tests. Field tests, which are trial runs of the test items, are designed to help ensure that the items function appropriately and are not confusing for students. This is typically accomplished by embedding field test items in the operational test. This is a commonly used and well-regarded practice that ensures the field test items are taken by a representative group of motivated students under standard conditions.

After the items have been field tested, another committee of Georgia educators examines the items again, along with the data from the field test. The committee reviews how the item performed in terms of how many students selected the correct answer and how many students selected each incorrect answer. The review also includes an analysis of how different groups of students performed to detect potential bias (i.e., Did the item appear to favor one group of students over another?). Once again the review committees have the authority to accept items as is, revise items for re-field testing, or reject items. Accepted items are then banked for future inclusion on an operational test form. Only after items have been field tested and approved by Georgia educators do they appear on an operational test form.

The next stage of test development consists of developing the actual test form that students will take. Items are carefully selected for a test form based on the blueprint developed by Georgia educators. Putting together a test form requires consideration of both content and statistical data. Each form of a test must assess the same range of content as well as carry the same statistical attributes.

When multiple test forms are used in a single administration or when a test is given in subsequent administrations (e.g., year to year tests), they must be equated. Equating refers to a statistical procedure to make sure that the tests are of equal difficulty. This is critical because it ensures that students are always held to the same standard. Additionally, it permits one to interpret differences in test performance as the result of changes in student achievement as opposed to fluctuations in the properties of the test form.

When a test is administered for the first time, standards must be established for the test. The standard setting process is the means by which educators decide what number of items a student must get correct (or how many total points) in order to meet or exceed expectations.

The final stage in test development is to produce scores and distribute results. Scores are typically reported as scale scores and performance levels. A scale score is based on the raw score (i.e., number of items correct or total points earned) on a test. The changing of raw score to scale scores is analogous to converting from the Centigrade scale to the Fahrenheit scale to report temperature. Scale scores are commonly used in large assessment programs. As an example, scores on the SAT, the widely used college entrance exam, are reported on a scale ranging from 200 – 800. Each time a new version of the SAT is administered, the raw scores are converted to this same scale in order to take into account any differences between various forms of the tests. Likewise, the Georgia Milestones results are presented in scale scores. This means that results can be consistently and meaningfully interpreted by students, parents, and educators. To assist the stakeholders in interpreting results, an interpretive guide is produced. This document sets forth how test scores should be interpreted and used and clearly indicates how scores for students with accommodations should be interpreted with caution.

By attending carefully to each phase of the test development process, the GaDOE can ensure that the Georgia Milestones Assessment System consists of valid instruments. The Georgia Milestones contractors produce documentation of each phase of the test development process and produce various pieces of evidence. The alignment of the Georgia Milestones assessments with the state's content standards and the reliance of input from Georgia educators at every phase of test development are critical to the test's validity. In addition, the department is collecting evidence through separate independent alignment studies to ensure that the test measures the state's content standards. The department will also be conducting over time analyses as evidence of external validity by comparing how the constructs measured by the Georgia Milestones assessments compare with other well-recognized assessments. The validation of a test is an ongoing process.

Reliability

For a test to be valid, it must also have reliability. However, the inverse of this is not true – a reliable measure is not necessarily valid. It is imperative that a test's validity must be established first and foremost, but aspects of the test's reliability must also be addressed during test development. So just what is reliability? Reliability is the degree to which test scores for a group of test takers are consistent and stable over time. In other words, a reliable assessment is one that would produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory effects.

For the Georgia Milestones Assessment System, Cronbach's alpha reliability coefficient (1951) is one reliability measure reported. A reliability coefficient expresses the consistency of test scores as the ratio of true score variance to observed total score variance (i.e., true score variance plus error variance). Cronbach's alpha measures the internal consistency over the responses to a set of items measuring an underlying unidimensional trait. Cronbach's alpha is computed using Crocker and Algina's formula (1986):

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right),$$

where k = number of items, σ_x^2 = the total score variance, and σ_i^2 = the variance of item i.

The reliability coefficient is a unitless index, which can be compared from test to test and ranges from 0 to 1. Tables 1 through 4 show the median reliability indices as well as the minimum and maximum values across forms and administrations for the Georgia Milestones assessments organized by subject area. These range from 0.85 to 0.94. The reliabilities are similar across grades/courses and subject areas and suggest that the Milestones assessments are sufficiently reliable for their intended purpose. That is, the reliability indicators obtained for Georgia Milestones suggest that scores reported to students for the 2015-2016 school year are well estimated and provide a reliable picture of student performance.

Table 1: Reliability Indicators for the 2015-2016 Georgia Milestones Assessments in English Language Arts

Grade/ Course	Number of Forms	Number of Operational Items per Form	Number of Raw Score Points per Form	Median Reliability	Minimum Reliability	Maximum Reliability
3	2	45	55	0.89	0.89	0.89
4	2	45	55	0.90	0.89	0.90
5	2	45	55	0.90	0.88	0.91
6	2	45	55	0.89	0.88	0.90
7	2	45	55	0.89	0.88	0.90
8	2	45	55	0.89	0.88	0.90
9th Grade Literature & Composition	4	45	55	0.88	0.87	0.89
American Literature & Composition	4	45	55	0.88	0.87	0.89

Table 2: Reliability Indicators for the 2015-2016 Georgia Milestones Assessments in Mathematics

Grade/ Course	Number of Forms	Number of Operational Items per Form	Number of Raw Score Points per Form	Median Reliability	Minimum Reliability	Maximum Reliability
3	2	53	58	0.92	0.92	0.92
4	2	53	58	0.92	0.91	0.93
5	2	53	58	0.93	0.92	0.93
6	2	53	58	0.92	0.92	0.92
7	2	53	58	0.93	0.92	0.93
8	2	53	58	0.91	0.90	0.91
Coordinate Algebra	4	53	58	0.90	0.89	0.93
Analytic Geometry	4	53	58	0.91	0.89	0.92
Algebra I	4	53	58	0.89	0.88	0.92
Geometry	4	53	58	0.91	0.91	0.94

Table 3: Reliability Indicators for 2015-2016 Georgia Milestones Assessments in Science

Grade	Number of Forms	Number of Operational Items per Form	Number of Raw Score Points per Form	Median Reliability	Minimum Reliability	Maximum Reliability
3	2	55	55	0.91	0.91	0.91
4	2	55	55	0.91	0.90	0.91
5	2	55	55	0.90	0.89	0.90
6	2	55	55	0.92	0.91	0.92
7	2	55	55	0.93	0.92	0.93
8	2	55	55	0.89	0.88	0.89
Biology	4	55	55	0.92	0.90	0.92
Physical Science	4	55	55	0.87	0.85	0.89

Table 4: Reliability Indicators for 2015-2016 Georgia Milestones Assessments in Social Studies

Grade	Number of Forms	Number of Operational Items per Form	Number of Raw Score Points per Form	Median Reliability	Minimum Reliability	Maximum Reliability
3	2	55	55	0.91	0.90	0.91
4	2	55	55	0.92	0.92	0.92
5	2	54	55	0.91	0.90	0.92
6	2	55	55	0.94	0.93	0.94
7	2	55	55	0.93	0.92	0.93
8	2	55	55	0.91	0.91	0.91
United States History	4	55	55	0.91	0.90	0.92
Economics/Business/ Free Enterprise	4	55	55	0.90	0.90	0.91

Summary

Foremost, the Georgia Milestones assessments have a high degree of validity because they serve the purpose for which they are intended – to measure student mastery of the state's content standards. Validity is established via the process of test development. The careful development from inception of the Georgia Milestones Assessment System and all steps in-between such as alignment with content standards, creation of test and item specifications, multiple reviews by educators, and careful form construction by content experts and psychometricians provide evidence that Georgia Milestones are valid instruments for the uses for which the department has developed the test. The reliability indices indicate that the tests provide consistent results and that the various generalizations of test results are justifiable. These strong indicators of reliability also support the tests' claim for validity.

References

To enhance one's understanding of the concepts presented in this brief, the following references are provided. The ones with an asterisk (*) are cited in the brief.

*American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Anastasi, A. (1988). *Psychological Testing*. New York: MacMillan.

*Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth.

*Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), 297-334.

*Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Hingham, MA: Kluwer-Nijhoff.

Hambleton, R. K. & Zaal, J.N., eds. (1991). *Advances in Educational and Psychological Testing*. Boston: Kluwer Academic.

Lyman, H. B. (1993). *Test Scores and What They Mean*. Boston: Allyn and Bacon.

McMillan, J. H. (2001). *Essential Assessment Concepts for Teachers and Administrators*. Thousand Oaks, CA: Corwin.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

Popham, W. J. (1998). *Classroom Assessment: What Teachers Need to Know*. Boston: Allyn and Bacon.

Thorndike, R. M. (1996). *Measurement and Evaluation in Psychology and Education* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

This brief is produced by Assessment Research and Development of the Georgia Department of Education.
Questions should be directed to the Assessment Research and Development staff at 404-656-2668.