WCEPS
WISCONSIN CENTER for
EDUCATION PRODUCTS & SERVICES

FINAL REPORT

Alignment Analysis of the ACT and SAT with the Georgia Standards of Excellence for American Literature and Composition, Algebra I, Geometry, and Biology

Sara C. Christopherson and Norman L. Webb

May 25, 2018

WebbAlign™

Wisconsin Center for Educational Products and Services
Matt Messinger, Executive Director
510 Charmany Drive, Suite 269
Madison, WI 53719

Acknowledgements

Algebra I:

External Panelists

Linda McQuillen	Group Leader	Wisconsin
Diane Briars		Pennsylvania
Michael Kestler		Washington, D.C

Georgia Panelists

Bobby Daniels	Decatur County, GA
Kelley Flournoy	Fayette County, GA
Meg Jett	Oconee County, GA
Melissa Schubert	Evans County, GA
Srinivasan Thiyagarajan	Richmond County, GA

Geometry:

External Panelists

Lynn Raith	Group Leader	Pennsylvania
Linda Hall		Washington, D.C
Jackie Snyder		Pennsylvania

Georgia Panelists

Wendy Dyer	Gwinnett County, GA
Mary Guy	Thomas County, GA
Leigh Moore	Laurens County, GA
Claire Sarver	Oconee County, GA
Michelle Taisee	Paulding County, GA

American Literature and Composition

External Panelists

Cindy Jacobson	Group Leader	Wisconsin
Greg Bartley		Wisconsin
Kymyona Burke		Mississippi

Georgia Panelists

Brandi Anthony	Toombs County, GA
Meshka Bailey	Forsyth County, GA
Christine Brand	Fayette County, GA
Kimberly Hernandez	Bibb County, GA
Alex Papanicolopoulos	Grady County, GA

Biology:

External Panelists

John Putnam	Group Leader	Virginia
Norman Dahm		Illinois
Jim Woodland		Nebraska

Georgia Panelists

Paula Cooper	Lumpkin County, GA
Mary-Melissa May	Gilmer County, GA
Theresa Senechek	Griffin-Spalding County, GA
Heather Toliver	Henry County, GA

The Georgia Department of Education, Atlanta, Georgia, funded this analysis. Dr. Allison Timberlake, Deputy Superintendent for Assessment and Accountability and Jonathan D. Rollins III, Measurement Program Manager for Assessment & Accountability, were the main contacts. Many other staff were also involved in the coordination of the alignment analysis.

Table of Contents

Table of Contents

For each content area:

Appendix A: Group Consensus DOK Values for Georgia Standards of Excellence

Appendix B: Data Analysis Tables for Each Test Form

Appendix C: Reviewers' Notes

Appendix D: Debriefing Summary Notes

Appendix E: Framework Analysis

Appendix F: DOK Definitions for Reading, Mathematics, and Science

Executive Summary

This report describes a two-stage alignment analysis conducted during the month of February, 2018, to provide information about the degree of alignment of the ACT and SAT with the Georgia Standards of Excellence (GSE). The content analysis was conducted to help inform a decision about whether or not school districts might be able to use either or both of these nationally-recognized college entrance tests in place of the Georgia Milestones End-of-Course assessments for American Literature and Composition, Algebra I, Geometry, and Biology. Evidence from this alignment study, along with evidence from other studies that the state of Georgia commissioned, will help the state to understand if the ACT and/or SAT could be used in lieu of the Georgia Milestones EOC assessments for fulfilling requirements as stated in Federal statute and Georgia legislation.

The alignment analysis consisted of two stages:

> **Stage I:** An analysis of ELA, mathematics, and science assessment framework documents; and
> **Stage II:** An in-person content alignment institute.

The first stage of the two-part alignment study compared the differences and similarities in the frameworks used to develop or interpret the findings from the ACT, SAT, and Georgia Milestones assessments. This information about the assessment structures and designs allowed for an analysis of convergent and divergent findings across the SAT and ACT when compared with the GSE with respect to the similarity of the constructs being measured. The ELA analysis was conducted by Dr. Erin Quast of Illinois State University, the mathematics analysis was conducted by Dr. Raven McCrory of Michigan State University, and the science analysis was conducted by Zoe Evans of Bowdon High School, Bowdon, Georgia. The reports from the framework analysis can be found in **Appendix E** for each subject area. The second stage of the analysis was a three-day in-person alignment institute that was held from February 12-14, 2018, in Atlanta, GA, to analyze the agreement between the Georgia Standards of Excellence for American Literature and Composition, Algebra I, Geometry and each of two forms of the ACT and the SAT and the Georgia Standards of Excellence for Biology and each of three forms of the ACT. Five Georgia educators and three external reviewers agreed to participate in each of the four subject-area analyses. Due to illness one Georgia biology educator was not able to participate. All panelists were selected because of their notable K-12 education experience and content expertise.

Overall, none of the test forms were found to be aligned with the GSE for any of the subjects. The ACT and SAT test forms had the greatest overlap with the GSE for American Literature and Composition and limited overlap with the standards for other courses. For American Literature and Composition, one of the ACT test forms was found to need slight adjustments—defined as needing six to 10 items revised or replaced—to meet the minimum cutoffs for full alignment. The other ACT test form was found to need major adjustments—defined as needing more than 10 items revised or replaced—to meet minimum alignment criteria. The ACT test forms reviewed would require approximately eight or approximately 16 items revised or replaced to meet minimum levels of acceptable alignment with the GSE for American Literature and Composition.

Both SAT test forms were found to need major adjustments to meet minimum levels of acceptable alignment with the GSE for American Literature and Composition, requiring approximately 13 or approximately 14 items revised or replaced.

The mathematics portions of both ACT and both SAT test forms analyzed would require major adjustments to meet minimum cutoffs for alignment with the corresponding GSE (for Algebra I or for Geometry). For the ACT test forms, only about 13% of items (8 of 60 items) or 23% of items (14 of 60 items) were judged by a majority of reviewers to correspond to an Algebra I standard. Only about 32% of ACT items (19 of 60 items) on each test form were judged by a majority of reviewers to correspond to a Geometry standard. For the SAT test forms, only about 62% of items (36 of 58 items) or 53% of items (31 of 58 items) were judged by a majority of reviewers to correspond to an Algebra I standard. Only about 16% of SAT items (9 of 58 items) on each test form corresponded to Geometry standards.

For Biology, none of the three ACT test forms were aligned with the GSE. Only 8%, 18%, or 20% of items corresponded to the GSE for Biology.

While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process and adds complexity to the administration of the tests, since items used to augment a test need to be administered separately from the college entrance test. Without such augmentation, however, these tests might not be viewed as meeting the United States Education Department (USED) criteria for aligned tests, thus jeopardizing the approval of the use of the college admissions tests in the federal requirements and the assessment peer review process.

Introduction and Methodology

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute solely of either of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). The corresponding methodology used to evaluate alignment has been refined and improved over the last 20 years, yielding a flexible, effective, and efficient analytical approach.

This is a report of a two-stage alignment analysis in the areas of American Literature and Composition, Algebra I, Geometry, and Biology that was conducted during the month of February, 2018, to provide information that could be used to judge the degree to which the ACT or SAT were aligned with the Georgia Standards of Excellence (GSE) used to develop the corresponding Georgia Milestones assessments. As such, the study focused on the degree to which the ACT and SAT test forms provided addressed the full depth and breadth of the GSE used to develop the Georgia Milestones assessments for American Literature and Composition, Algebra I, Geometry, and Biology.

The alignment analysis consisted of two stages:

> **Stage I:** An analysis of ELA, mathematics, and science assessment framework documents; and
> **Stage II:** An in-person content alignment institute.

The Stage I framework analysis for ELA was conducted by Dr. Erin Quast of Illinois State University, the framework analysis for mathematics was conducted by Dr. Raven McCrory of Michigan State University, and the framework analysis for science was conducted by Zoe Evans of Bowdon High School, Bowdon, Georgia. Each subject area education expert analyzed the specification of content in supporting documents for each of the ACT, SAT, and Georgia Milestones, including blueprints, item specifications, item type, and other relevant materials that were used in developing tests or interpreting scores. The framework analysis yielded a comparison of overall test claims and assessment targets, descriptions of how specific terms and concepts were used in each of the frameworks, and identification of any relevant structural variation among the three frameworks for each content area including any differences in item types, emphasis in content topics, type of reading passages used, sizes of numbers used, and other factors. Contextual factors such as the allotted time for essay writing were also considered. Full reports from the framework analysis are included in **Appendix E** of this report for each subject area. Findings from the framework analyses are also summarized in the Findings section of this report.

The Stage II in-person content alignment institute was held over three days, February 12-14, in Atlanta, GA, at the Courtyard by Marriott Atlanta Decatur Downtown/Emory. The ELA and mathematics portions of two test forms of each of the ACT and SAT were reviewed at the institute. Three test forms of the ACT science test were also reviewed. Eight reviewers served on each of the ELA, Algebra I, and Geometry panels. Seven reviewers served on the Biology panel; one Georgia panelist was not able to attend due to illness. An experienced group leader facilitated each panel. Study director Norman Webb is the researcher who developed the alignment study procedures and criteria (through the National Institute for Science Education in 1997, funded by the National Science Foundation, and in cooperation with the Council of Chief State School Officers) that influenced the specification of alignment criteria by the U.S. Department of Education. The Webb alignment process has been used to analyze curriculum standards and assessments in at least 30 states to satisfy or to prepare to satisfy Title I compliance as required by the United States Department of Education (USED). Study Technical Director Sara Christopherson has participated in and led Webb alignment studies since 2005 for state departments of education as well as for other entities.

The Version 2 of the Web Alignment Tool (WATv2) was used to enter all of the content analysis codes during the institute. The WATv2 is a web-based tool connected to the server at the Wisconsin Center for Education Research (WCER) at the University of Wisconsin-Madison. It was designed to be used with the Webb process for analyzing the alignment between assessments and standards. Prior to the institute, a group number was set up on the WATv2 for each of the four panels. Each panel was assigned one or more group identification numbers and the group leader was designated. Then the reporting categories and standards were entered into the WATv2 along with the information for each assessment, including the number of items, the weight (point value) given to each item, and additional comments such as the identification number for the item to help panelists find the correct item. A sequential account of the alignment study procedures is provided below.

Training and Coding

In the morning of the first day of the alignment institute, reviewers in all four content area groups received an overview of the purpose of their work, the coding process, and general training on the Depth-of-Knowledge (DOK) definitions used to describe content complexity. All reviewers had some understanding of the DOK levels prior to the institute. The general training at the alignment institute was crafted to contextualize the origins of DOK (to inform alignment studies of standards and assessments) and purpose (to differentiate between and among degrees of complexity), and to highlight common misinterpretations and misconceptions to help reviewers better understand and, therefore, consistently apply the depth of knowledge (DOK) language system. Panelists also practiced assigning DOK to sample assessment items that were selected to foster important discussions that promote improved conceptual understanding of DOK. Appropriate training of the panelists at the alignment institute is critical to the success of the project. A necessary outcome of training is for panelists to have a common, calibrated understanding of the DOK language system for describing categories of complexity.

The groups were then separated into different rooms to receive more detailed training on the DOK levels for each content area. Through interactive and participatory training, panelists reviewed the content area-specific definitions of the four DOK levels and worked toward a common understanding of the difference between and among each of the levels of complexity. Because the two mathematics groups used the same DOK definitions, they completed this portion of the training together, to promote consistency between the two groups' use of DOK as it pertains to mathematics. Definitions for each DOK level for ELA, mathematics, and science are included within this report. Reviewers then worked to calibrate their use of DOK to evaluate the complexity of a subset of the standards, first assigning DOK individually and then participating in a consensus discussion. After completing coding and discussion of the subset, the panelists reviewed the DOK levels previously assigned to the standards, when available (completed by other expert panels using a similar process) and flagged any standards that they wanted to discuss further, that they thought needed clarification, and/or that had a DOK assigned that they thought should be considered for adjustment because it did not accurately depict the appropriate level of content complexity. Group leaders facilitated discussions for any standards that one or more panelists flagged. If the discussion resulted in a decision to change the DOK that was assigned to a standard, then that change was made in the online data collection system, the WATv2. This study included all standards identified by Georgia that defined the expectations for the corresponding high school courses: American Literature and Composition, Algebra I, Geometry, and Biology.

The Georgia Standards of Excellence for American Literature and Composition, Algebra I, and Geometry were derived from the Common Core State Standards (CCSS) and can therefore be considered as meeting the requirement of high quality standards related to college and career readiness. The Georgia Standards of Excellence for Biology are grounded in Project 2061's *Benchmarks for Science Literacy* (1993) and the National Research Council's *A Framework for K-12 Science Education* (2012). These conceptual frameworks for science education are intended to prepare students to be scientifically literate adults, prepared to pursue post-secondary education and/or careers in the sciences. As such, the GSE for Biology can also be considered as meeting the requirement of high quality standards related to college and career readiness.

After thoroughly discussing the standards and coming to consensus on the intended complexity of each standard, panelists then conducted individual analyses of 3-5 assessment items from the first ACT test form and the first SAT test form (for ELA and mathematics groups). For each item, panelists worked individually to assign a DOK level to the item and then to code each item to the standard that they judged the item to measure, i.e. what students are expected to know or do in order to respond to the question. Up to three standards could be coded as corresponding to each item.

Following individual analyses of the items, reviewers participated in a debriefing discussion in which they analyzed the degree to which they had coded particular items or types of content to the standards. This overall process was repeated at the start of each test form to maintain calibration within each group of reviewers. Reviewers then completed analysis of the remaining items individually for each test form.

As reviewers work, they become increasingly familiar with the standards. They also refine their approach to interpretation and analysis of content. To ensure that the novice effect would be equally distributed across both the ACT and SAT test forms, half of the ELA and math groups' panelists coded the ACT first and half of the groups' panelists coded the SAT first for each test form.

Reviewers were instructed to focus primarily on the alignment between the GSE and the assessment items on the ACT and SAT test forms. However, reviewers were encouraged to offer their opinions on the standards or on the assessment tasks by writing a note about the item in the appropriate text box in the WATv2 data collection tool. Reviewers were instructed to enter a note into the WATv2 for an assessment item if the item only corresponded to a part of a standard and not the full standard. Thus, the reviewers' notes can be used to reveal if assessment items only targeted a part of the individual standards. Reviewers also could indicate whether there was a Source-of-Challenge issue with an item—i.e. a technical problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly. No Source-of-Challenge issues were identified on any of the assessments.

Reviewers engaged in adjudication of their results after completing the coding of each test form. After discussing an item, the reviewers were given the option to make changes to their codings, but were not required to make any changes if they thought their coding was appropriate. After all of the reviewers completed coding an assessment form, the study director and group leader identified the assessment items that did not have a majority of reviewers in agreement on DOK or where the reviewers differed significantly on the DOK assigned (e.g. three different DOK values were assigned). When these substantial disagreements occur, it suggests that reviewers are either interpreting the DOK definitions in very different ways or are interpreting the particular assessment item in very different ways.

Reviewers also discussed items for which there were great differences in coding to a standard. The adjudication process helped panelists identify and correct any errors in coding (e.g. accidentally assigning an item to a standard that they did not intend to assign). Adjudication also helped panelists build familiarity with the standards (e.g. a reviewer might not have noticed that a particular expectation is explicit in one of the standards) as well as build common interpretation of the standards (e.g. panelists may calibrate their understanding of the meaning of certain standards that may be interpreted in different ways due to ambiguous wording or due to differences in the way people understand the content). Adjudication also helped reveal differences in interpretation of assessment items, and helped reviewers to build a common understanding of exactly what content particular items were assessing. Overall, adjudication is intended to foster full and appropriate interpretation of the assessment items and standards, and to ensure that panelists have coded their items as they intended. Reviewers were not required to change their results after the discussion. Reviewer agreement statistics were computed after adjudication and are included in the Findings section of this report.

Reviewers were instructed to consider the full statement of expectations to consider if an assessment item should be mapped to a standard. In some cases, reviewers could make reasonable arguments for coding an item to different standards. For example, both ELAGSE11-12RL4 and ELAGSE11-12L4.a include the expectation that students use context clues to identify the meaning of unknown words and phrases.

If reviewers map an item to a variety of standards it may also indicate that the assessment task may be inferred to relate to more than one standard but that the item is not a close match. Reviewers may have difficulty finding where an item best fits when an assessment is coded to a set of standards that were not used in developing the assessment. If an item did not closely fit any standard, then the reviewers were instructed to code the item to a standard where there was a partial, but reasonable, fit or to a conceptual category level: the strand level for ELA GSE standards or domain level for mathematics GSE. Coding to the level of a conceptual category may be referred to as coding to a "generic" standard.

All seven biology reviewers coded all ACT science test forms and the biology group adjudicated after completing each ACT test form. Math and ELA groups adjudicated after the first ACT and SAT forms were completed and then again after the second ACT and SAT forms were completed. Mathematics and ELA reviewers were working at different paces within their respective groups, and several reviewers were only able to complete three of the four test forms assigned. By the end of the time allotted for coding, eight ELA reviewers coded ACT form 74C and seven ELA reviewers completed coding of ACT form A10. Eight ELA reviewers coded each of the two SAT test forms (April and October 2017). Seven algebra reviewers coded each of ACT form 74C and form A10. Eight algebra reviewers completed coding SAT form April 2017 and seven algebra reviewers completed coding SAT form October 2017. Seven geometry reviewers completed SAT test form October 2017 and all eight reviewers competed the other three test forms.

Data Analysis

To derive the results from the analysis, the reviewers' responses were averaged. First, the value for each of the four alignment criteria (described in the next section) was computed for each individual reviewer. Then the final reported value for each criterion was found by averaging the values across all reviewers. Any variance among reviewers was considered legitimate, for example, with the reported DOK level for an item falling somewhere between the two or more assigned values. Such variation could signify differences in interpretation of an item or of the assessed content and/or a DOK that falls in between two of the four defined levels. Any large variations among reviewers in the final results represented true differences in opinion among the reviewers and were not because of coding error. These differences could be due to different standards targeting the same content knowledge or may be because an item did not explicitly correspond to any standard, but could be inferred to relate to more than one standard. Standard deviations are reported in the tables provided in **Appendix B**, which give one indication of the variance among reviewers.

The results produced from the institute pertain only to the issue of alignment between the Georgia Standards of Excellence and the nine assessments that were analyzed. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met.

Alignment Criteria Used for This Analysis

This report describes the results of an alignment study of nine test forms or portions of test forms (ELA, mathematics, science) with the corresponding GSE. The study addressed specific criteria related to the content agreement between the standards and assessments. Four criteria received major attention:

- Categorical Concurrence,
- Depth-of-Knowledge Consistency,
- Range-of-Knowledge Correspondence, and
- Balance of Representation.

Details on the criteria and indices used for determining the degree of alignment between standards and assessments are provided below. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had reasonably met the expectations within the reporting categories for each discipline. In the descriptions below, the words "domain" and "reporting category" are used to describe reporting levels.

Reporting Categories and Standards:

Study results are reported according to the reporting categories (RCs) for each content area. These RCs are given below. For each content group, reviewers individually assigned DOK to a subset of standards and then engaged in a consensus discussion to promote group calibration of DOK use as well as to foster deep understanding of the standards. Previously assigned DOK levels were reviewed for the remaining ELA and mathematics standards, with thorough discussion of any standards for which one or more reviewers proposed adjustment or requested further consideration. If the group chose to adjust any of the previously assigned DOKs, these adjusted consensus DOK levels were entered into the WATv2 for use in the study. The biology panel assigned a DOK to each biology standard and then participated in a consensus discussion to reconcile any differences in codings. These consensus values were then entered into the WATv2. Consensus DOK values for all standards are given in **Appendix A** for each subject.

In this analysis, the reporting categories for **ELA** were:

- Reading literary (RL)
- Reading informational (RI)
- Writing (W)
- Language (L)

Total number of standards: 65

The reporting categories for **Algebra I** were:

- The Real Number System (N.RN)
- Quantities (N.Q)
- Seeing Structure in Expressions (A.SSE)
- Arithmetic with Polynomials & Rational Expressions (A.APR)
- Creating Equations (A.CED)
- Reasoning with Equations and Inequalities (A.REI)
- Interpreting Functions (F.IF)
- Building Functions (F.BF)
- Linear, Quadratic, and Exponential Models (F.LE)
- Interpreting Categorical and Quantitative Data (S.ID)

Total number of standards: 60

The reporting categories for **Geometry** were:

- Congruence (G.CO)
- Similarity, Right Triangles, and Trigonometry (G.SRT)
- Circles (G.C)
- Expressing Geometric Properties with Equations (G.GPE)
- Geometric Measurement and Dimension (G.GMD)
- Modeling with Geometry (G.MG)
- Conditional Probability and the Rules of Probability (S.CP)

Total number of standards: 45

The reporting categories for **Biology** were:

- GSE.SB1. Obtain, evaluate, and communicate information to analyze the nature of the relationships between structures and functions in living cells.
- GSE.SB2. Obtain, evaluate, and communicate information to analyze how genetic information is expressed in cells.
- GSE.SB3. Obtain, evaluate, and communicate information to analyze how biological traits are passed on to successive generations.
- GSE.SB4. Obtain, evaluate, and communicate information to illustrate the organization of interacting systems within single-celled and multi-celled organisms.
- GSE.SB5. Obtain, evaluate, and communicate information to assess the interdependence of all organisms on one another and their environment.
- GSE.SB6. Obtain, evaluate, and communicate information to assess the theory of evolution.

Total number of standards: 24

Mapping of Items to Standards:
If no particular grade-level standard was targeted by a given assessment item, reviewers were instructed to code the item at a "higher" or more inclusive level, such as the, strand level for ELA, or domain level for mathematics. This coding to a "generic standard" sometimes indicates that the item was inappropriate for a particular grade level (for example, the item might better match a standard from another grade level). If the item was grade-appropriate but an appropriate standard was not found, a generic coding may indicate that there is a part of the content within the standards that is being interpreted differently by different parties. Generic coding may also occur when mapping a test to a set of standards that is different from the set used to develop the test. In this case, some items on an assessment may simply target a different set of learning expectations, as would be expected per framework analyses findings.

In the descriptions below, the term "standards" may be used as an umbrella term, to refer to expectations in general. In addition to judging alignment between reporting categories and assessments on the basis of the four key alignment criteria, reviewers had the opportunity to identify and comment on any items with Source-of-Challenge and other issues.

Categorical Concurrence
An important aspect of alignment between standards and assessments is whether both address the same content categories. The Categorical-Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of Categorical Concurrence between standard and assessments is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each reporting category. The analysis assumed that the assessment had to have at least six items (or points for polytomous items) for measuring content from a reporting category in order for a minimum acceptable level of Categorical Concurrence to exist between the domain and the assessment. The number of items/points, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is 0.1, it was estimated that six items would produce an agreement coefficient of at least 0.63. This indicates that about 63% of the group would be consistently determined to be masters or non-masters if two equivalent test administrations were employed. The agreement coefficient would increase to 0.77 if the cutoff score is increased to one standard deviation from the mean and, with a cutoff score of 1.5 standard deviations from the mean, to 0.90.

Usually states do not report student results by domains or require students to achieve a specified cutoff score on expectations related to a domain. If a state did do this, then the state would seek a higher agreement coefficient than 0.63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a reporting category, and as a basis for making some decisions about students' knowledge of that content under the reporting category. If the mean for six items is 3.0 points and one standard deviation is equal to a one-point item, then a cutoff score set at 4.0 points

would produce an agreement coefficient of 0.77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. Depth-of-Knowledge Consistency between standards and an assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards. For consistency to exist between the assessment and the reporting categories, as judged in this analysis, at least 50% of the items corresponding to a reporting category had to be at or above the depth-of-knowledge level of the corresponding content expectation. The 50% level, a conservative minimum cutoff point, is based on the assumption that a minimal passing score for any one reporting category of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the content expectations within the corresponding reporting categories. For example, assume an assessment included six items related to one domain and students were required to answer correctly four of those items to be judged proficient—i.e. 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one expectation. If a domain had between 40% and 50% of items at or above the depth-of-knowledge levels of the expectations, then it was reported that the criterion was “weakly” met.

DOK Levels

Interpreting and assigning depth-of-knowledge levels to both standards and assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent for reading, mathematics, and science.

DOK Levels for Reading

DOK 1

DOK 1 involves reading text orally and with basic comprehension, decoding words, blending phonemes, receiving and reciting facts, demonstrating letter and word knowledge, and recognizing text features and common spelling patterns. DOK 1 also includes receiving or reciting facts acquired by processing text as well as reading orally without the analysis of text. Very basic comprehension of a text gained from knowledge of vocabulary and explicit structure of the text is at this category. Tasks require only a shallow understanding of the text presented and often consist of verbatim recall from text, slight paraphrasing of specific details from the text, or simple understanding of a single word or phrase. Younger students who answer direct questions about features stated explicitly in the text are performing at this category. Applying phonics and word analysis skills in decoding words are also DOK 1 tasks.

Some examples that represent, but do not constitute all of, DOK 1 performance include:

- Support ideas with reference to verbatim (or only slightly paraphrased) details from the text.
- Use a dictionary to find the meanings of words.
- Recognize figurative language in a reading passage.

DOK 2

DOK 2 involves drawing meaning from text by using organizational structure, evidence, and context; summarizing main ideas, character traits, plots, themes, and figurative use of words; following cause-effect sequences and multiple ideas through a text; distinguishing among hypotheses and givens as well as fact from opinion; and explaining differences among genres (poetry, expository materials, fiction, etc.). DOK 2 requires the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis or inference is required. DOK 2 tasks may require use of specific information from the text to explain given events and ideas. At this level, reading concepts (e.g. making inferences or predictions) are generally applied for purposeful reading. Multiple features of the text are processed to gain a deeper understanding of the text such as organizing in a time sequence, outlining, comparing fact from opinion, and using graphic aides. Deciphering main ideas supported by key details or drawing on details to describe a feature in a story are stressed. Younger students conveying important points from a story fit under this category. DOK 2 ideas, in general, apply the skills and concepts that constitute DOK 1. However, DOK 2 activities involve closer understanding of text, possibly through paraphrasing, such as putting in one's own words both the question and response to an assessment item. Some examples that represent, but do not constitute all of, DOK 2 performance include:

- Use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

DOK 3

DOK 3 involves conducting analyses of the text to make inferences about author's purpose and use of textual features (e.g. literary devices to support and convey the main message); engaging in critical reading to attest to the credibility of the message, the internal logic, and implied values, attitudes, and biases; and going beyond the text by comparing features and meaning with other texts, considering the impact of the time period and other conditions when the text was written, and raising valid alternative hypotheses and conclusions to those presented in the text. At DOK 3, deep knowledge becomes a greater focus. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas while applying reasoning and planning. Students must be able to support their thinking. Younger students who provide some valid evidence for their breakdown of a story into meaningful parts are performing at this category.

Tasks at a Category 3 may involve abstract theme identification, inference across an entire passage with multiple paragraphs, or students' application of prior knowledge. Activities may also involve identifying more abstract connections between texts. Some examples that represent, but do not constitute all of, DOK 3 performance include:

- Explain or recognize how the author's purpose affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

DOK 4

DOK 4 involves at least as complex content as in the previous category, but also requires working on a task over an extended period of time such as when conducting a research project over a period of weeks. The extended time that accompanies this type of activity allows for creation of original work and requires metacognitive awareness that typically increases the complexity of a DOK 4 task overall, in comparison with DOK 3 activities. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking.

DOK 4 activities may have students take information from multiple passages and texts to find supporting evidence and counter points for developing an argument or reaching conclusions or could involve creating an original thesis on a topic based on information drawn from relevant references. For younger students, an extended period of time could be multiple days for reaching conclusions from reading a number of texts. Students take information from a multiple of passages and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts requiring work over an extended period of time. Some examples that represent, but do not constitute all of, DOK 4 performance include:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

DOK Levels for Mathematics

DOK 1 (Recall)

DOK 1 is defined by the rote recall of information or performance of a simple, routine procedure. For example, repeating a memorized fact, definition, or term, performing a simple algorithm, rounding a number, or applying a formula are DOK 1 performances. Performing a one-step computation or operation, executing a well-defined multi-step procedure or a direct computational algorithm are also included in this category. Examples of well-defined multi-step procedures include finding the mean or median or performing long division. Reading information directly from a graph, plugging data into an electronic device to derive an answer, or simple paraphrasing are all tasks that are considered a level of complexity comparable to recall. A student answering a Level 1 item either knows the answer or does not: that is, the item does not need to be "figured out" or "solved."

At a DOK 1, problems in context are straightforward and the solution path is obvious. For example, the problem may contain a keyword that indicates the operation needed. Other DOK 1 examples include plotting points on a coordinate system, using coordinates with the distance formula, or drawing lines of symmetry of geometric figures.
At more advanced levels of mathematics, symbol manipulation and solving a quadratic equation or a system of two linear equations with two unknowns are considered comparable to recall assuming students are expected or likely to use well-known procedures (e.g. factoring, completing the square, substitution, or elimination) to derive a solution. Operating on polynomials or radicals, using the laws of exponents, or simplifying rational expressions are considered rote procedures.
Verbs should not be classified as any category without considering what the verb is acting upon or the verb's direct object. "*Identify* attributes of a polygon" is recall, but "*identify* the rate of change for an exponential function" requires a more complex analysis. To *describe* by listing the steps used to solve a problem is recall (i.e. *Show your work*) whereas to *describe* by providing a mathematical argument or rationale for a solution is more complex.

DOK 2 (Skill/Concept)
DOK 2 involves engaging in some mental processing beyond a habitual response as well as decision-making about how to approach the problem or activity. This category can require conceptual understanding and/or demonstrating conceptual knowledge by explaining thinking in terms of concepts.
DOK 2 tasks includes distinguishing among mathematical ideas, processing information about the underlying structure, drawing relationships among ideas, deciding among and performing appropriate skills, applying properties or conventions within a relevant and necessary context, transforming among different representations, interpreting and solving problems and/or graphs. When given a problem statement, formulating an equation or inequality, deriving a solution, and reporting the solution in the context of the problem fit within DOK 2. Processes such as classifying, organizing, and estimating that involve attending to multiple attributes, features, or properties also fall into this category. Verifying that the number of objects in one set is larger or fewer than the number of objects in a second set by matching pairs or forming equivalent groups is a DOK 2 activity for a kindergartener. A first grader modeling a joining or separating situation pictorially or physically also is in this category.
Skills and concepts include constructing a graph and interpreting the meaning of critical features of a function, beyond just identifying or finding such features as well as describing the effects of parameter changes. Note, however, that using a well-defined procedure to find features of a standard function, such as the slope of a linear function with one variable or a quadratic, is a DOK 1. Graphing higher order or irregular functions is a DOK 2. Basic computation, as well as converting between different units of measurement, are generally a Category 1, but illustrating a computation by different representations (e.g. equations and a base-ten model) to explain the results is a DOK 2. Computing measures of central tendency (applying set procedures) is a DOK 1, but interpreting such measures for a data set within its context or using measures to compare multiple data sets is a DOK 2. Performing original formal proofs is beyond DOK 2, but explaining in one's own words the reasons for an action or application of a property is comparable to a DOK 2. Activities at a DOK 2 are not limited only to number skills, but may involve visualization skills (e.g. mentally rotating a 3D figure or transforming a figure) and probability skills requiring more than simple counting (e.g.

determining a sample space or probability of a compound event). Other activities at this category include detecting or describing non-trivial patterns, explaining the purpose and use of experimental procedures, and carrying out experimental procedures.

DOK 3 (Strategic Thinking)

DOK 3 requires reasoning and analyzing using mathematical principles, ideas, structure, and practices. DOK 3 includes solving involved problems; conjecturing; creating novel solutions and forms of representation; devising original proofs, mathematical arguments, and critiques of arguments; constructing mathematical models; and forming robust inferences and predictions. Although DOK 2 also involves some problem solving, DOK 3 includes situations that are non-routine, more demanding, more abstract, and more complex than DOK 2. Such activities are characterized by producing sound and valid mathematical arguments when solving problems, verifying answers, developing a proof, or drawing inferences. Note that the sophistication of a mathematical argument that would be considered DOK 3 depends on the prior knowledge and experiences of the person. For example, primary school student arguments for number problems can be a DOK 3 activity (e.g. counting number of combinations, finding shortest route from home to school, computing with large numbers) as can abstract reasoning in developing a logical argument by students in higher grades. DOK 3 problems are those for which it is not evident from the first reading what is needed to derive a solution and so require demanding reasoning to work through. Such problems usually can be solved in different ways and may even have more than one correct solution based on different stated assumptions. Paraphrasing in one's own words or reproducing a proof that was previously demonstrated is a DOK 2. Applying properties and producing arguments in proving a theorem or identity not previously seen is a DOK 3. Also in the DOK 3 category is making sense of the mathematics in a situation, creating a mathematical model of a situation considering contextual constraints, deriving a new formula, designing and conducting an experiment, and interpreting findings.

DOK 4 (Extended Thinking)

DOK 4 demands are at least as complex as those of DOK 3, but a main factor that distinguishes the two categories is the need to perform activities over days and weeks (DOK 4) rather than in one sitting (DOK 3). The extended time that accompanies this type of activity allows for creation of original work and requires metacognitive awareness that typically increases the complexity of a DOK 4 task overall, in comparison with DOK 3 activities. Category 4 activities require complex reasoning, planning, research, and verification of work. Conducting a research project, performance activity, an experiment, and a design project as well as creating a new theorem and proof fit under Category 4. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, collecting water temperature from a river each day for a month and then reporting the findings by constructing a graph is a DOK 2 activity. Developing a mathematical model of the flow of water in a river for all four seasons using a number of variables would be a DOK 4 activity. It is likely that a DOK 4 activity will require making connections among a number of ideas or variables within the area of mathematics or among a number of content areas. Category 4 activities require selecting an appropriate approach among many alternatives to produce a product, conclusion, or finding, such as critiquing a body of work, synthesizing ideas in a new way, or creating an original model.

DOK Levels for Science

DOK 1 (Recall and Reproduction)

DOK 1 is defined by the recall of information, such as a fact, definition, or term, as well as performance of a simple grade-level-appropriate science process or procedure. DOK 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Simple word problems that can be directly translated into and solved by a formula are considered DOK 1.

A student answering a DOK 1 item either knows the answer or does not: that is, the item does not need to be "figured out" or "solved." In other words, if the knowledge necessary to answer an item automatically provides the answer to it, then the item is at DOK 1.

Some examples that represent, but do not constitute all of, DOK 1 performance are:

- Recall or recognize a fact, term, structure, or property.
- Represent in words or diagrams a scientific concept or relationship.
- Provide or recognize a standard scientific representation for simple phenomenon.
- Perform a grade level-appropriate routine procedure, such as measuring length or completing a basic Punnett square.

Verbs such as "identify," "recall," "recognize," "use," "calculate," and "measure" generally represent cognitive work at the recall and reproduction level. Verbs such as "describe" and "explain" could be classified at different DOK levels, depending on the complexity of what is to be described and explained. Note, however, that verbs should not be the basis of DOK classification without considering what the verb is acting upon or the verb's direct object.

DOK 2 (Skills and Concepts)

DOK 2 includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in DOK 1. Items require students to make some decisions about how to approach the question or problem. Classifying and comparing are activities that are typically a DOK 2 as well as organizing and displaying data in tables, graphs, and charts. These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomena and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different DOK levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a DOK 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at DOK 3.

Some examples that represent, but do not constitute all of, DOK 2 performance, are:

- Specify and explain the relationship between facts, terms, properties, or variables.
- Describe and explain examples and non-examples of science concepts.
- Select a procedure according to specified criteria and perform it.
- Formulate a routine problem, given data and conditions.
- Organize, represent, and interpret data.
- Interpret or explain phenomena in terms of science concepts.
- Make basic predictions for cause-and-effect relationships.

DOK 3 (Strategic Thinking)
DOK 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at DOK 3 are complex and abstract. The complexity does not result only from the fact that there could be multiple answers, a possibility for both DOK 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to provide a rationale for their thinking is at DOK 3 (although a task requiring a very simple explanation or a word or two should be at DOK 2). An activity that has more than one possible answer and requires students to justify the response they give would most likely be a DOK 3. Experimental designs at DOK 3 may involve more than one dependent variable. Some examples that represent, but do not constitute all of DOK 3 performance, are:

- Identify research questions and design investigations for a scientific problem.
- Use concepts to solve non-routine problems.
- Draw robust conclusions from observations.
- Cite evidence and develop a logical argument.
- Develop a scientific model for a complex situation.
- Form conclusions from experimental data.

DOK 4 (Extended Thinking)
DOK 4 demands are at least as complex as those of DOK 3, but a main factor that distinguishes the two categories is the need to perform activities over days and weeks (DOK 4) rather than in one sitting (DOK 3). The extended time that accompanies this type of activity allows for creation of original work and requires metacognitive awareness that typically increases the complexity of a DOK 4 task overall, in comparison with DOK 3 activities. On-demand assessment instruments very rarely include assessment activities that could be classified as DOK 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. "Develop generalizations of the results obtained and the strategies used and apply them to new problem situations," is an example of a grade 8 objective that is a DOK 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought over extended time will be DOK 4.

Note that the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. DOK 4 requires complex reasoning, experimental design and planning, as well as an extended period of time for completion. For example, if a student

has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a DOK 2 activity. However, if the student designs and conducts a river study that involves all aspects of a scientific investigation, from forming a testable question to communication of results, this would be a DOK 4. Some examples that represent, but do not constitute all of, a DOK 4 performance are:

- Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.
- Analyze the results of multiple studies on a particular science topic to form an original conclusion about the subject.
- Evaluate strengths and weaknesses of an experimental design and develop a revised experimental design.

Range-of-Knowledge Correspondence

For reporting categories and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The Range-of-Knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a reporting category and an assessment considers the number of standards within the reporting category with one related assessment item/activity. Fifty percent of the standards for a reporting category must have at least one related assessment item for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a reporting category. This assumes that each expectation for a reporting category should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for a reporting category increases the likelihood that students will have to demonstrate knowledge on more than one expectation per reporting category to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectations that can be assessed. Range-of-Knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of reporting categories and a large number of expectations. If 50% or more of the objectives for a reporting category had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If between 40% and 50% of the objectives for a reporting category had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned reporting categories and assessments require that knowledge be distributed equally or proportionally in both. The Range-of-Knowledge criterion only considers the number of expectations with at least one assessment item within a reporting category; it does not take into consideration how the assessment items/activities are distributed among these expectations. *The Balance-of-Representation criterion is used to indicate the degree to which one standard*

is given more emphasis on the assessment than another. An index is used to judge the distribution of assessment items. This index only considers the expectations for a reporting category that has at least one related assessment item per expectation. The index is computed by considering the difference in the proportion of expectations and the proportion of items assigned to the expectation. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a reporting category are equally distributed among the expectations for the given reporting category. Index values that approach 0.0 signify that a large proportion of the items assess only one or two of all of the expectations that were measured. Depending on the number of expectations and the number of items, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining expectations) has an index value of less than 0.5. A bimodal distribution has an index value of around 0.55 or 0.6. Index values of 0.7 or higher indicate that items/activities are distributed among all of the expectations at least to some degree (e.g. nearly every expectation has at least two items) and is used as the acceptable level on this criterion. Index values between 0.6 and 0.7 indicate the Balance-of-Representation criterion has only been "weakly" met.

Source-of-Challenge Criterion

The Source-of-Challenge criterion is used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted language reporting category or expectation (i.e. construct irrelevance). Bias and sensitivity issues as well as technical issues and error could all be reasons for an item to have a Source-of-Challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed. (No items were flagged with Source-of-Challenge in this study.)

Cutoffs for Alignment Criteria

For overall alignment, an assessment form is reported as *fully aligned* if no items need replacement to meet the conditions for all of the criteria described above. Note that "fully aligned" refers to the condition of meeting the *minimum acceptable levels* of alignment and does not mean that an assessment has "100% alignment" with the corresponding standards. A test form is considered *acceptably aligned* if it needs between one and five items replaced or revised in order to meet the minimum acceptable conditions for all alignment criteria. A test form is reported to *need slight adjustments* if six to ten items need to be replaced or revised to meet the minimum levels of alignment criteria and is reported to *need major adjustments* if more than ten items need to be replaced or revised. These categories represent typically used cutoff levels.

Findings: American Literature and Composition

Framework Analysis for ELA

Dr. Quast's framework analysis for ELA assessments mapped convergent and divergent aspects of the Georgia American Literature and Composition EOC, the ACT, and the SAT (see **Appendix E** for ELA for full framework analysis). All three tests assess student knowledge and skills on aspects of reading, language, and writing. However, notable differences exist across assessments in terms of assessment structure, student writing, reading passages, and content. Differences in test structure include variations in item type(s), allotted test time, and number of test items/tasks.

Content differences reflect the divergent assessment targets (GSE, ACT CCR Standards, SAT Skills). The framework analysis found that the ACT College and Career Readiness Standards had a close match to just 26% of the GSE for American Literature and Composition and a partial match to 38% of the standards. The SAT Skills had a close match to only 40% of the GSE for American Literature and Composition and a partial match to 54% of the standards. In other words, around half of the GSE for American Literature and Composition do not correspond to any of the ACT CCR Standards or SAT Skills. Both the ACT and SAT include assessment targets outside of the content within the GSE, such as identification of subject-verb agreement, pronoun-antecedent agreement, inappropriate shifts in verb tense and other expectations. Text complexity was described differently for each assessment, preventing a direct comparison. The Georgia EOC design is such that the passages used on the assessment are intended to reflect the specifications in the corresponding standards. ACT test forms include passages with a range of text complexity. SAT test forms include text that is "complex." Additional detail is provided in **Appendix E**.

A comparison of session times, item counts, and item types are provided in **Table 1**. While the ACT and SAT each included multiple choice items and one essay, the Georgia EOC includes multiple choice items as well as technology enhanced, constructed response, extended constructed response, and extended writing response items.

Table 1. Georgia EOC, ACT, & SAT Item Counts, Types, and Session Times - ELA

	Test Sections & Time	**Total Number of Items**	**Item Type** (across sections)
Georgia American Literature and Composition EOC	Section 1 (writing) *90mins* Section 2 *75mins* Section 3 *75mins* **Total 240min**	**Total:** 60 items	Selected Response Technology-enhanced Constructed Response (CR) Extended CR Extended Writing Response
ACT English, Reading, & Writing	English *45min* Reading *35min* Writing *40min* **Total 120min**	75 items 40 items **T: 115** 1 essay	Multiple-choice
SAT Reading, English and Language, & Essay	Reading *65min* Writing/Language *35min* Essay *50min* **Total 150min**	52 items 44 items **T: 96** 1 essay	Multiple-choice

Source: Georgia Department of Education, 2017; The College Board, 2015, ACT, 2014

As shown in **Table 1** above, the ACT had the most items (115 + 1 essay), SAT had slightly fewer (96 + 1 essay), and the Georgia EOC had the fewest number of items (60, including writing prompts). Because the Georgia EOC test form contained fewer test items than the ACT or SAT but were administered in longer sessions, students would have significantly more time to work each item than they would on the corresponding ACT and SAT tests. Average time per non-essay item/task is shown in **Table 2** below.

Table 2. Time per assessment item/task for Georgia EOC, ACT, and SAT ELA tests – excluding essay

Test	**Number of Items**	**Assessment Time**	**Average Time per Item***
Georgia EOC (Sections 2 + 3)	55	150 min	2.7 min
ACT English and Reading	115	80 min	0.7 min
SAT Reading and Writing/Language	96	100 min	1.0 min

*Note that some items may take more time than others.

Table 3 provides an overview of the content and context between the Georgia EOC Section 1 (essay), ACT essay, and the SAT essay. A significant difference was the amount of time allocated to student writing, with the Georgia EOC allowing nearly three

times more time than the ACT allowed and nearly twice the time that the SAT allowed. The nature of the essays was also slightly different: the Georgia assessment task includes three MC items and one CR item prior to the extended writing prompt. These items are intended to support students in making sense of the provided passage. The Georgia EOC essay may be argumentative, informative, or explanatory, while the ACT task required an argumentative essay, and the SAT task required a written analysis of a source text.

Table 3. Comparison of Georgia EOC, ACT, & SAT Essays

	Georgia Writing Essay	**ACT Writing Essay (optional)**	**SAT Writing Essay (optional)**
Time	90 minutes	35 minutes	50 minutes
Item Format	3 MC items; 1 CR item; 1 extended constructed response (essay)	Stimulus (issue description & two texts providing two different viewpoints) & 1 prompt	Single source text & 1 prompt
Essay Type	Argumentative, Informative, or Explanatory	Argumentative	Written Analysis of Source text
Rubric Domains	1. Idea development, organization, and coherence 2. Language usage and conventions	1. Ideas & Analysis 2. Development & Support 3. Organization 4. Language Use	1. Reading 2. Analysis 3. Writing

There were no field test items on the ELA portions of the ACT and SAT test forms and no items were excluded from the ELA analysis. On all test forms, all items except for the writing prompt were weighted as one point. The ACT essay was weighted at 12 points, total, reflective of a total subject-level writing score of up to 12, based on four domain scores. The four domain scores correspond to a four-part rubric scored on a scale of 2-12 and then averaged to yield the subject-level score. The SAT essay was weighted at 24 points, reflective of three scores, corresponding to a three-part rubric, each scored on a scale of 2-8. Scores are reported for each of the three rubric dimensions.

Standards

A summary of the levels of complexity within the GSE for American Literature and Composition is given in **Table 4**. Eight of the standards included in the study (12%) were considered DOK 1. These expectations were all within the Language domain, and targeted conventions of Standard English as well as basic use of reference materials. Twenty standards (31%) were considered a DOK level 2, emphasizing work that involves both comprehension and subsequent processing of text, as well as making basic inferences from text and using specific information from text to explain events and ideas. The largest group of standards, twenty-nine standards (45%), were considered to be DOK 3, emphasizing expectations for deep analysis of text and abstract thinking, including making holistic inferences based on text, and engaging in critical reading to consider aspects of author's purpose and use of textual features. Eight standards (12% percent) were considered DOK 4. A DOK 4 expectation is one that is both at least as

complex as a DOK 3 but also requires extended time—days, weeks, or months—to complete. Although some components of these DOK 4 standards may be reasonably assessed by on-demand assessments, DOK 4 standards should not be expected to be fully assessed by an on-demand test. All of the expectations used in this study will be referred to as *standards,* because they were all the equivalent unit of analysis, although some statements of expectations were standards while others were elements. Elements are subparts of standards and are designated by a letter (a, b, or c).

Table 4. Expectations by Depth-of-Knowledge (DOK) Levels for GSE for American Literature and Composition, February, 2018

ELA	Total Number of Expectations	DOK Level	Number of Standards by Level	Percent within RC by Level
ELAGSE11-12RL READING LITERARY	9	2 3	4 5	44 56
ELAGSE11-12RI READING INFORMATIONAL	10	2 3 4	3 6 1	30 60 10
ELAGSE11-12W WRITING	28	2 3 4	5 16 7	18 57 25
ELAGSE11-12L LANGUAGE	18	1 2 3	8 8 2	44 44 11
Total	65	1 2 3 4	8 20 29 8	12 31 45 12

Mapping of Items by Standards

There were no items on either test form of the ACT or SAT that a majority of reviewers coded to a generic standard.

The ACT test forms included slightly more items (115 items) than the SAT test forms (96 items) but targeted a slightly lower percentage of the total GSE for American Literature and Composition than the SAT (see **Table 5**). Averaging across the two ACT test forms and across the two SAT test forms, the ACT forms were found to include items that addressed around 30% of the course GSE compared with around 36% of the GSE for the items on the the SAT test forms.

Table 5. Number and Percent of GSE for American Literature and Composition with at least One Corresponding Item Found by a Majority of Reviewers

Assessment	Number of Items (including writing prompt)	Number of GSE Targeted	Percentage of Total GSE with at least One Corresponding Assessment Item
ACT Form 74C	116	19	29%
ACT Form A10	116	20	31%
SAT Apr 2017	97	24	37%
SAT Oct 2017	97	23	35%

Comparison of Overall DOK Distribution

A comparison of the overall DOK distribution for the multiple choice portions of each assessment, averaged across the two test forms, is shown in **Table 6**. The essay portions of both the ACT and the SAT were considered DOK 3 and are excluded from the calculation. Considering DOK 2 and DOK 3 items as "higher complexity," the SAT contained a slightly greater proportion of higher-DOK items (79%) compared with the ACT (66%).

On the ACT, the distribution of the items by DOK levels varies some from the intended distribution of DOK levels for the ACT forms, according to the ACT Technical Manual. The English test blueprint specifies 33-41% DOK 3 items and the Reading test blueprint specifies 25-50% DOK 3 items while reviewers coded only 7% of all items as DOK 3. This suggests that the DOK definitions used by ACT are different from the original Webb definitions, which have been updated and were used in this study. The intended DOK definitions for SAT were not provided.

Table 6. DOK Distribution, averaged on two test forms for ACT and SAT ELA portions, essay excluded

Test	DOK 1	DOK 2	DOK 3
ACT	34%	59%	7%
SAT	21%	71%	8%

Alignment Statistics and Findings for ACT and SAT Test Forms and GSE for American Literature and Composition

Overall alignment results are summarized in **Table 7** below and then detailed for each test form in the pages that follow. The main alignment issue for both the ACT forms and the SAT forms was Range-of-Knowledge, with the ACT test forms not meeting this criterion or only weakly meeting this criterion for all four reporting categories and the SAT test forms not meeting this criterion or only weakly meeting this criterion for three of the four reporting categories. One ACT form was found to need slight adjustments (defined as 6-10 items revised or replaced) to meet minimum cutoffs for alignment while

the other needed major adjustments (defined as more than 10 items revised or replaced). Both SAT test forms were found to need major adjustments (more than 10 items revised or replaced). to meet minimum cutoffs for alignment. These findings are shown in **Table 7**.

Table 7. Overall Alignment Findings for Two Forms Each of ACT and SAT ELA Assessments with GSE for American Literature and Composition

Test Form	Alignment Findings	Approximate Number of Items that Need Revision/Replacement for Full Alignment
ACT 74C	Needs Slight Adjustments	8
ACT A10	Needs Major Adjustments	16
SAT April 2017	Needs Major Adjustments	13
SAT Oct 2017	Needs Major Adjustments	14

Results by Test Form

The results of the analysis for each of the four alignment criteria are provided in **Tables 8 to 11** for each ELA test form for Reporting Categories RL, RI, W, and L. The approximate numbers of replaced or revised items necessary to meet minimum levels of alignment are provided for each test form. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form. The reviewers' notes and debriefing comments (**Appendices C** and **D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In **Tables 8 to 11**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth-of-Knowledge Consistency, 10% under an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation. Categorical Concurrence is reported in average number of items. Depth of Knowledge Consistency is reported by the percent of items that were at or above the DOK of the corresponding standard. Range-of-Knowledge is reported as the percent of standards within each reporting category that were targeted by one or more items. Balance of representation is an index value, ranging from 0-1.

ACT Test Forms

For ACT Form 74C, three items would need to be revised or replaced to meet the Range-of-Knowledge criterion for the Reading Literary (RL) reporting category. If these items targeted RL standards that were not yet targeted, the weakness in Balance of Representation could also be resolved. For the Reading Informational (RI) reporting category, one item would need to be revised or replaced to meet the Range-of-Knowledge criterion. For the Writing (W) reporting category, one item would need to be revised or replaced to meet the DOK Consistency criterion. If this item also targeted a writing standard that is not currently assessed, then only one more item would need to be revised or replaced to meet the Range-of-Knowledge criterion. For the Language (L) reporting category, about two items would need to be revised or replaced to meet DOK Consistency. Overall, for ACT Form 74C, a total of approximately eight items would need to be revised or replaced to meet the minimum levels of acceptable alignment.

Table 8. ACT Form 74C June 2017 – ELA with GSE for American Literature and Composition

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literary (RL)	10	69%	30%	0.61	YES	YES	NO	WEAK
Reading Informational (RI)	30	70%	41%	0.70	YES	YES	WEAK	YES
Writing (W)	38	49%	42%	0.72	YES	WEAK	WEAK	YES
Language (L)	61	59%	36%	0.58	YES	YES	NO	NO

*Number of items

For ACT Form A10, approximately three items would need to be revised or replaced to meet the Range-of-Knowledge criterion for the Reading Literary (RL) reporting category. If these items targeted RL standards that were not yet targeted, the weakness in Balance of Representation could also be resolved. For the Reading Informational (RI) reporting category, one item would need to be revised or replaced to meet the Range-of-Knowledge criterion. For the Writing (W) reporting category, one item would need to be revised or replaced to meet the DOK Consistency criterion. If this item also targeted a W standard that is not currently assessed, then an additional four items would need to be revised or replaced to meet the Range-of-Knowledge criterion. For the Language (L) reporting category, two items would need to be revised or replaced to meet DOK Consistency. If these two items targeted L standards that were not yet targeted, then an additional five items would need to be revised or replaced to also target standards that are not yet targeted to meet the Range-of-Knowledge criterion. Overall, for ACT Form A10, a total of approximately 16 items would need to be revised or replaced to meet the minimum levels of acceptable alignment.

Table 9. ACT Form A10 December 2017 – ELA with GSE for American Literature and Composition

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literary (RL)	10	79%	30%	0.57	YES	YES	NO	NO
Reading Informational (RI)	28	66%	43%	0.60	YES	YES	WEAK	WEAK
Writing (W)	28	49%	34%	0.78	YES	WEAK	NO	YES
Language (L)	55	41%	15%	0.86	YES	WEAK	NO	YES

*Number of items

SAT Test Forms

For SAT Form April 2017, two items would need to be revised or replaced to meet the Range-of-Knowledge criterion for the Reading Literary (RL) reporting category. For the Reading Informational (RI) reporting category, standard RI.1 is more strongly emphasized than the other RI standards. Twenty two percent of items on the test form target this standard. If this emphasis is acceptable for Georgia, then it would not be considered as an alignment issue because the other three alignment criteria are met for the RI reporting category. If this degree of emphasis on RI.1 is *not* acceptable to the state, then Balance could be improved with the addition of five to ten items distributed among the other standards within the RI reporting category. Alternatively, Balance could also be improved by the removal of items from the assessment. For the Writing (W) reporting category, six items would need to be revised or replaced that targeted W standards that are not currently assessed to meet the Range-of-Knowledge criterion. For the Language (L) reporting category, one item would need to be revised or replaced to meet DOK Consistency. If this item targeted a L standard that is not currently assessed, then four additional items would need to be revised or replaced to meet Range-of-Knowledge. Overall, for SAT Form April 2017, a total of approximately 13 items would need to be revised or replaced to meet the minimum levels of acceptable alignment.

Table 10. SAT Form April 2017 – ELA with GSE for American Literature and Composition

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literary (RL)	10	62%	35%	0.83	YES	YES	NO	YES
Reading Informational (RI)	66	83%	76%	0.59	YES	YES	YES	NO
Writing (W)	39	65%	29%	0.57	YES	YES	NO	NO
Language (L)	30	48%	26%	0.72	YES	WEAK	NO	YES

*Number of items

For SAT Form October 2017, two items would need to be revised or replaced to meet the Range-of-Knowledge criterion for the Reading Literary (RL) reporting category. For the Reading Informational (RI) reporting category, the weak Balance of Representation is typically not considered an alignment issue because the other three alignment criteria are met. As with SAT Form April 2017, if the emphasis on RI.1 is not acceptable to the state, Balance could be improved by the addition or removal of items. For the Writing (W) reporting category, six items would need to be revised or replaced that targeted Writing standards that are not currently assessed to meet the Range-of-Knowledge criterion. For the Language (L) reporting category, two items would need to be revised or replaced to meet DOK Consistency. If these items targeted a standard that is not currently assessed, then four additional items would need to be revised or replaced to meet Range-of-Knowledge. Overall, for SAT Form October 2017, a total of approximately 14 items would need to be revised or replaced to meet the minimum levels of acceptable alignment.

Table 11. SAT Form October 2017 – ELA with GSE for American Literature and Composition

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
Reading Literary (RL)	10	69%	43%	0.73	YES	YES	WEAK	YES
Reading Informational (RI)	68	81%	75%	0.61	YES	YES	YES	WEAK
Writing (W)	40	68%	30%	0.57	YES	YES	NO	NO
Language (L)	30	39%	18%	0.82	YES	NO	NO	YES

*Number of items

Writing Prompts Each assessment included a single weighted writing prompt that was evaluated according to a three-part or four-part rubric. The 35-minute ACT essay is argumentative, addressing W.1.a, W.1.b, W.1.c, and W.1.d (which relate to the writing of arguments to support claims) as well as L.1 and L.3, related to the use of language. The 50-minute SAT essay was coded to W.2 and W.2.b (which relate to the writing of informative or explanatory texts) as well as RI.1, RI.2, and RI.3 which relate to gathering information or evidence from other texts, integrating the information, and using it to support an analysis. This reflects the structure of the SAT essay, which is a written analysis of source text.

Source of Challenge Issues and Reviewers' Comments Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have about an item. A Source-of-Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items for which more than one reviewer left a Source-of-Challenge comment.

Reviewers also wrote notes about many items on each form. Some notes indicate when only part of a particular standard was targeted by an assessment task. These notes also include general comments as well as indicate concerns with items. Some notes include suggestions for resolutions to issues identified. After coding each assessment form, reviewers were asked to respond to four debriefing questions. The full text of reviewers' notes and debriefing comments can be found in **Appendices C** and **D**.

Comments: ACT Test Forms Multiple reviewers left comments related to a perception that the level of sophistication of questions on the ACT were low compared with the expectations of the Georgia Standards of Excellence, expressing that basic comprehension was overemphasized and that not enough items engage students in interpretation of texts or higher order thinking about one or more texts at a time. More

than one reviewer expressed some mismatch between the ACT Form 74C reading passages and the Georgia American Literature and Composition expectations. One reviewer noted "the absence of any documents of historical and literary significance on the ACT. Such documents are central to 11th grade American Literature curricula throughout the state of Georgia." More than one reviewer preferred the passages used in ACT Form A10 compared with Form 74C, noting they were "more relatable," "fine," and "appropriate for [the] subject matter." One reviewer noted that there was "a good variety of Language and Writing questions," although another reviewer commented that items on ACT Form 74C were directed at "simply retaining grammar rules" while SAT items "seemed to have more questions regarding writing choices."

Comments: SAT Test Forms Multiple reviewers left comments about the SAT test forms' focus on reading analysis skills and informational text and that there was less of a focus on literature, which contrasts with the core emphasis of the American Literature and Composition course. Reviewers' perspectives differed on the passages, for example, reviewers commended the "quality and range of texts," noted that "there is a good variety of passages," and that they were "strong" and "engaging," but one reviewer commented that the reading selections were "overly difficult…wordy, and not motivating to read." One reviewer noted that, in contrast to the ACT, "foundational U.S. documents of historical significance do appear on the SAT." This same reviewer, however, did not think that either the ACT or SAT test forms adequately addressed the GSE, and commented that "both exams appear to be biased toward fiction, and neglect other literary genres covered in the curriculum for American Literature (grade 11). This would be a particularly grave concern for Georgia teachers who teach American Literature."

Reliability among Reviewers

Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 12**, on the following page, were computed after adjudication. The overall intraclass correlation among the ELA reviewers' assignment of DOK levels to items was high (0.91 or higher) for all analyses (**Table 12**). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for every item the coding assigned by each reviewer compared to the coding by each of the other seven reviewers. For example, for eight reviewers a total of 28 comparisons were computed for each item. For most alignment studies, the standards pairwise agreement is higher than 0.6. The pairwise agreement for assigning standards to items was reasonably high for all test forms. Some "decision rules" helped to guide agreement, to ensure consistent interpretation of item types and standards. For example, if a multiple choice item required a student to write by proxy (putting the student in the role of the writer) then reviewers agreed it would be appropriate to correlate to a writing standard, even if the student was not actually writing. For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For all test forms, pairwise agreement for reporting category is reasonably high or high.

Table 12. Intraclass and Pairwise Comparisons, ACT, and SAT with GSE for American Literature and Composition

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
ACT 74C	0.94	0.64	0.92	0.66
ACT A10	0.91	0.70	0.96	0.80
SAT April 2017	0.95	0.75	0.95	0.73
SAT Oct 2017	0.95	0.78	0.96	0.73

Summary of Comparisons of the Two Assessments

A summary of alignment results by test form is provided in **Table 13**. The two ACT test forms were found to need slight or major adjustments in order meet minimum cutoffs for alignment with the GSE. Both SAT test forms were found to need major adjustment to meet minimum cutoffs for alignment with the GSE.

Table 13. Percent of GSE Reporting Categories for American Literature and Composition with Acceptable Level on Each Alignment Criteria when Compared to Four Test Forms

Assessment Form	Categorical Concurrence (Percent of RCs with over six items)	Depth-of-Knowledge Consistency (50% at/above)	Range-of-Knowledge (50% of standards)	Balance of Representation (without possible weakness)
ACT 74C	100%	75%	0%	50%
ACT A10	100%	50%	0%	50%
SAT Apr 2017	100%	75%	25%	50%
SAT Oct 2017	100%	75%	25%	50%

Findings: Algebra

Framework Analysis for Mathematics – Algebra I

Professor Raven McCrory, a mathematics educator at Michigan State University, conducted a review of the design documents and other explanatory materials found for each of the three assessments. This report is included as **Appendix E** for Algebra I. Information from this report was used to compare the GSE for Algebra I and Geometry to content expectations as specified in design documents from the ACT and the College Board. This analysis also provided information on the structure of the ACT and SAT and administration instructions for both. The design documents included test blueprints, test specifications, and curriculum standards as were available.

About 20 of the 45 Algebra I standards (44%) did not have comparable standards in any of the documents found for the ACT or SAT assessments. For example, the Algebra I standards include standards related to students understanding exponents, radicals, and rational and irrational numbers (RC3: N-RN1.1-1.2). These content topics were not found in the SAT materials. These topics were found in ACT College and Career Readiness Standards, but are not among the benchmarks considered at the level of college and career readiness.

Another difference among the frameworks was in the area of statistics and probability. The Georgia Algebra I standards included standards RC3: S-ID.3.8 and 9 (computation and interpretation of correlation coefficients for a linear line of best fit). Neither of the SAT or ACT documents reviewed in the framework analysis considered this topic as an essential understanding for college and career readiness. Also, some differences were found in the description of items. For example, the Georgia Algebra I assessment specifications explicitly noted that items written for certain standards should be embedded in a problem context. No such explicit statements were found for the ACT or SAT. Thus, the framework analysis did reveal some design differences and variation in the content intended to be assessed. Additional detail is provided in **Appendix E**.

A comparison of session times, item counts, and item types are provided in **Table 14**. The mathematics assessments differed in their structure and the type(s) of items. The ACT mathematics assessment consisted of 60 items completed in 60 minutes, all items are equally weighted at one point. All 60 items were multiple choice with five choices. Calculators were permitted for use when taking the ACT mathematics test but not required. Students could use most calculators, including four-function, scientific, or graphing calculators except for those explicitly prohibited such as those with built-in or downloaded algebra computer system functionality.

The SAT mathematics assessment had 58 items administered in two parts, including 20 items where calculators were not permitted and 38 items where students were permitted to use a calculator. All items were equally weighted at one point. The College Board SAT website provides a list of brands and models of calculators that are acceptable for use on the mathematics test. Permitted calculators include most graphing calculators and all scientific calculators. More basic four-function calculators are permitted but not recommended. Students were allotted 80 minutes to complete the mathematics proportion of the assessment. The SAT assessments had two types of items, multiple choice (78%) and grid-ins (22%), in which students fill in a grid to enter a positive whole number, decimal, or fraction (**Table 14**). The Georgia Algebra I EOC has 73 total items including a variety of item types. Of these, 52 items (for a total of 58 points) contribute to a student's final score, and the selected response items are all equally weighted at one point.

Table 14. Georgia EOC, ACT, & SAT Item Types–Mathematics

Test	Item Type										
	Multiple-choice		Constructed Response (CR)		Extended CR		Technology-enhanced		Fill-in-the-grid		Total Number
	N	%	N	%	N	%	N	%	N	%	
Algebra I EOC	69	95	2	3	1	1	1	1	--	--	73
ACT	60	100	--	--	--	--	--	--	--	--	60
SAT	45	78	--	--	--	--	--	--	13	22	58

Table 15. Time per assessment item/task for Georgia EOC, ACT, and SAT–Mathematics

Test	Number of Items	Assessment Time	Average Time per Item**
Algebra I EOC	73 items*	170 min	2.3 min
ACT Form A10	60 items	60 min	1 min
SAT Apr 2017	58 items	80 min	1.4 min

*Of these, 10 items are field test items and only some of the Norm-Referenced Test items (those that correspond to GSE) will contribute to a student's Criterion-Referenced score; total item count was included to calculate average time per item.
**Note that some items may take more time than others.

Standards

A total of 60 expectations were used in this Algebra I study. All of these expectations will be referred to as *standards,* because they were all an equivalent unit of analysis used in this study, even though 45 of these statements of expectations are standards (e.g. MGSE9-12.F.IF.8) and 15 are elements (e.g. MGSE9-12.F.IF.8a). Elements are subparts of standards and are designated by a letter (a, b, or c). Seven standards had 1-3 elements for a total of 15 elements. The standards and elements were organized into 10 domains which are considered the reporting category for the study. Reviewers were asked to review assigned DOKs to standards used in previous studies. They were asked to change any DOK they disagreed with and to assign a DOK to any element that did not have a DOK. A consensus process was used. For the Algebra I study, the reviewers changed the DOK of seven standards by making the level higher (DOK 1 to 2 or DOK 2 to 3) and decreased the DOK on two standards (DOK 2 to 1). **Table 16** summarizes the DOKs assigned to the GSE Algebra I standards—15 DOK 1 and 45 DOK 2. Most of the standards were judged to require students to demonstrate conceptual understanding or procedural knowledge. One-fourth of the standards required students to demonstrate recall of information.

Table 16. Percent of Expectations by Depth-of-Knowledge (DOK) Levels for the Mathematics Georgia Standards of Excellence for Algebra I

Domain	Total Number of Expectations	DOK Level	Number of Standards by Level	Percent within RC by Level
N.RN The Real Number System	2	1 2	1 1	50 50
N.Q Quantities	3	2	3	100
A.SSE Seeing Structure in Expressions	7	1 2	1 6	14 86
A.APR Arithmetic with Polynomials & Rational Expressions	1	1	1	100
A.CED Creating Equations	4	1 2	1 3	25 75
A.REI Reasoning with Equations & Inequalities	10	1 2	4 6	40 60
F.IF Interpreting Functions	12	1 2	2 10	17 83
F.BF Building Functions	4	2	4	100
F.LE Linear, Quadratic, and Exponential Models	7	1 2	3 4	43 57
S.ID Interpreting Categorical & Quantitative Data	10	1 2	2 8	20 80
Total	60	1 2	15 45	25 75

Mapping of Items by Standards

There were no items on either test form of the ACT or SAT that a majority of reviewers coded to a generic standard. Neither assessment had items that targeted a high percentage of the GSE for Algebra I. At most, SAT Form April 2017 had items that corresponded to about one-third of the Georgia GSE (**Table 17**).

Table 17. Number and Percent of Mathematics GSE for Algebra I with at least One Corresponding Item Found by a Majority of Reviewers

Test	Total Number of Items	Number of GSE Targeted	Number of Algebra I Standards with Corresponding Item(s)
ACT Form 74C	60	11	18%
ACT Form A10	60	8	13%
SAT Apr 2017	58	21	36%
SAT Oct2017	58	15	26%

Table 18. Number and Percent of Mathematics Items for ACT and SAT Assessments Judged by Majority of Reviewers as Corresponding to Mathematics GSE for Algebra I Standards

Test	Total Items	Items Corresponding to Algebra I Standards		Indecisive		Items Corresponding to Other Mathematics Content	
		Number	Percent	Number	Percent	Number	Percent
ACT Form 74C	60	13	22%	--	--	47	78%
ACT Form A10	60	10	17%	5	8%	45	75%
SAT Apr 2017	58	35	60%	1	2%	22	38%
SAT Oct 2017	58	30	52%	1	2%	27	46%

Although both assessments had nearly the same number of items (58 or 60 items; **Table 18)**, the assessments varied in the proportion of items that targeted Algebra I standards. Across both ACT test forms, a majority of reviewers found that around 20% of the total items mapped to at least one of the 60 GSE Algebra I standards. Across both SAT test forms, a majority of reviewers found a greater proportion of items, 56%, that mapped to at least one GSE Algebra I standard.

Comparison of Overall DOK Distribution

A comparison of the overall DOK distribution for items on each assessment, averaged across the two test forms, is shown in **Table 19**. Both assessments had nearly three-quarters of the Algebra I items with a DOK level 2, 70% for the ACT and 78% for the SAT. The remaining items were judged to have a DOK level 1. No items that mapped to a GSE Algebra I standard were judged to have a DOK level 3 by a majority of the reviewers. DOK 3 items are not necessary per Algebra I GSE; none of the 60 standards were considered a DOK 3 level expectation.

Table 19. DOK Distribution of Algebra I Items, averaged across two test forms, for the ACT and SAT

Test	DOK 1	DOK 2	DOK 3
ACT	30%	70%	0%
SAT	22%	78%	0%

Alignment Statistics and Findings for ACT and SAT Test Forms and GSE for Algebra I

Overall alignment results are summarized in **Table 20** below and then detailed for each test form in the pages that follow. A main alignment issue for both the ACT forms was a lack of Categorical Concurrence for any of the reporting categories. An associated main alignment issue was a corresponding unmet Range-of-Knowledge for all but one to two of the ten reporting categories for the ACT test forms. Although the SAT forms had a

greater proportion of items mapped to Algebra I standards, the test forms still had unmet Categorical Concurrence for eight to nine out of ten of the reporting categories and an associated unmet Range-of-Knowledge. Both ACT and both SAT test forms were found to need major adjustments to meet minimum cutoffs for alignment (**Table 20**).

Table 20. Overall Alignment Findings for Two Forms Each of ACT and SAT Mathematics Assessments with GSE for Algebra I

Test Form	Alignment Findings	Number of Items that Need Revision/Replacement for Minimum Alignment
ACT 74C	Needs Major Adjustments	48
ACT A10	Needs Major Adjustments	48
SAT April 2017	Needs Major Adjustments	33
SAT Oct 2017	Needs Major Adjustments	35

Results by Test Form

The results of the analysis for each of the four alignment criteria are provided in **Tables 21** to **24** for each mathematics test form for the ten Reporting Categories. The approximate numbers of replaced or revised items necessary to meet minimum levels of alignment are provided for each test form. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form. The reviewers' notes and debriefing comments (**Appendices C** and **D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In **Tables 21** to **24**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth-of-Knowledge Consistency, 10% under an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation. Categorical Concurrence is reported in average number of items. Depth of Knowledge Consistency is reported by the percent of items that were at or above the DOK of the corresponding standard. Range-of-Knowledge is reported as the percent of standards within each reporting category that were targeted by one or more items. Balance of representation is an index value, ranging from 0-1. Alignment statistics for DOK Consistency, Range-of-Knowledge, and Balance are reported only for reporting categories that have three or more corresponding items.

The GSE structure includes ten reporting categories, corresponding to ten domains. A test form that met the typical minimum levels for alignment with a set of Algebra I standards that includes ten reporting categories would need a minimum of 60 items, with all items corresponding to Algebra I, at an appropriate level of DOK, and targeting at least half of the standards within each domain. If fewer reporting categories were used by grouping the domains into larger categories, alignment results would be affected and, potentially, improved for all test forms.

ACT Test Forms

Reviewers found 10 items (Form A10) or 13 items (Form 74C) on the ACT forms that clearly mapped to GSE for Algebra I standards. On Form A10, reviewers were inconclusive on five other items. These items may have had some relation to Algebra I, but were not directly related to any of the standards. Of the 45 or 47 items that were judged to map to a content area other than to one of the Algebra I standards, 15 or 17 items (about 25%) were judged to be geometry items, five or six items (about 10%) were judged to correspond to probability, about 15 or 16 items (about 25%) were judged to correspond to the general area of number, and nine items (15%) corresponded to content areas beyond Algebra I, such as log equations or trigonometry. About 33% of the items on both the ACT mathematics forms addressed content related to middle school mathematics and 15% corresponded to mathematics topics generally taught after Algebra I.

For both of the ACT mathematics forms that were analyzed in the study, less than a quarter of the items mapped to the GSE for Algebra I standards. Each form had at least one item that mapped to five or six of the 10 domains. There was some variation between the two forms, but most of the Algebra I items mapped to four of the 10 Algebra I domains—Creating Equations (CED), Reasoning with Equations and Inequalities (REI), Interpreting Functions (IF), and Interpreting Categorical and Quantitative Data (ID). On each ACT form from two to four items mapped to standards under each of these four domains. In most alignment studies, six items for a domain or reporting category is considered the minimum number of items needed to make some reliable judgment about a student's proficiency. Neither ACT form met this minimum cutoff for any of the 10 domains.

The items that did correspond to the GSE for Algebra I compared favorably to the level of complexity as expected by the corresponding standards. For most of the domains by each ACT form, the Depth-of-Knowledge Consistency criterion was acceptably met, with 50% or more of the items with a DOK level that was the same or higher than the DOK level of the corresponding standard. Minimum cutoffs for DOK Consistency were met for all domains by the ACT Form 74C and for all but two domains by the ACT Form A10. In general, the items on the two ACT forms required the level complexity as expected by the corresponding standards.

The range of items among the standards under a domain was low in part because of too few items corresponded to each domain. Only one domain, A.CED, acceptably met the Range-of- Knowledge Correspondence criteria on each of the two ACT forms. The A.CED (Creating Equations) domain had four underlying standards. For each form at least two items mapped to two different standards under that domain. Range was acceptably met for the A.APR (Arithmetic with Polynomials and Rational Expressions)

domain with Form 74C because that domain only had one underlying standard. Reviewers found one item on Form 74C that mapped to this one standard. For the remainder of the domains, the two forms did not have a sufficient number of items adequately distributed among the standards to say that the assessments covered the breadth of content as expressed by the GSE for Algebra I.

Because of the low number of items for each domain, Balance of Representation does not have very much meaning. The Balance Index is computed only using the standards with at least one corresponding item. There were not enough items distributed among the standards under a domain to yield useful information about the degree of emphasis among the standards by the items.

Overall, both ACT forms, 74C and A10, would have to be supplemented by approximately 48 items to be considered fully aligned to the GSE for Algebra I. These items would need to be selected carefully to have at least six items for each Algebra I domain. If there were fewer domains by aggregating the existing ones into larger clusters, then fewer items would be needed to include in a supplemental assessment. The current corresponding items, assuming similar items would be chosen to supplement the current 13 items, suggests that the items would have an appropriate DOK level. Carefully selecting the items to target specific standards would also ensure that range would successfully be met along with balance. However, without being supplemented, the two ACT forms used in the analysis cannot be considered as aligned to the GSE for Algebra I.

Table 21. ACT Form 74C June 2017 – mathematics with GSE for Algebra I

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
N.RN The Real Number System	1	--	--	--	NO	N/A	N/A	N/A
N.Q Quantities	0	--	--	--	NO	N/A	N/A	N/A
A.SSE Seeing Structure in Expressions	0	--	--	--	NO	N/A	N/A	N/A
A.APR Arithmetic with Polynomials & Rational Expressions	1	--	--	--	NO	N/A	N/A	N/A
A.CED Creating Equations	2	--	--	--	NO	N/A	N/A	N/A
A.REI Reasoning with Equations & Inequalities	2	--	--	--	NO	N/A	N/A	N/A
F.IF Interpreting Functions	3	93%	18%	0.87	NO	YES	NO	YES
F.BF Building Functions	0	--	--	--	NO	N/A	N/A	N/A
F.LE Linear, Quadratic, and Exponential Models	0	--	--	--	NO	N/A	N/A	N/A
S.ID Interpreting Categorial & Quantitative Data	3	93%	21%	0.85	NO	YES	NO	YES

*Number of items

Table 22. ACT Form A10 June 2017 – Mathematics with GSE for Algebra I

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
N.RN The Real Number System	0	--	--	--	NO	N/A	N/A	N/A
N.Q Quantities	0	--	--	--	NO	N/A	N/A	N/A
A.SSE Seeing Structure in Expressions	0	--	--	--	NO	N/A	N/A	N/A
A.APR Arithmetic with Polynomials & Rational Expressions	0	--	--	--	NO	N/A	N/A	N/A
A.CED Creating Equations	4	72%	56%	0.87	NO	YES	YES	YES
A.REI Reasoning with Equations & Inequalities	3	94%	27%	0.84	NO	YES	NO	YES
F.IF Interpreting Functions	2	--	--	--	NO	N/A	N/A	N/A
F.BF Building Functions	1	--	--	--	NO	N/A	N/A	N/A
F.LE Linear, Quadratic, and Exponential Models	0	--	--	--	NO	N/A	N/A	N/A
S.ID Interpreting Categorical & Quantitative Data	2	--	--	--	NO	N/A	N/A	N/A

*Number of items

SAT Test Forms

Over half of the items on both SAT forms were judged to correspond to GSE for Algebra I. Reviewers found 35 items (60%) of the items on SAT Form April 2017 and 30 items (52%) of the items on SAT Form October 2017 that mapped to Algebra I standards. Based on their notes (Appendix C), the other items on the April 2017 Form mapped to Geometry (8 items; 14%), Number (6 items; 10%), Probability/Statistics (1 item; 2%), and more advanced topics than Algebra I (7 items; 12%). In addition to Algebra I, on SAT Form October 2017 reviewers found items that they described as Geometry (6 items; 10%), Number (7 items; 12%), Probability/Statistics (5 items; 9%), and more advanced topics than Algebra I (9 items; 16%). Of these items, reviewers noted that seven items on SAT Form April 2017 and eight items on SAT Form October 2017, or about 12-15% of the total of 58 items, corresponded to mathematics content taught in middle schools.

The majority of reviewers found at least one item on each of the two SAT forms that mapped to nine of the ten GSE domains for Algebra I. This indicates that the two SAT forms had some breadth in content coverage. The SAT Form April 2017 had four or more items that mapped to five of the domains (**Table 23**). The SAT Form October 2017 had four or more items that mapped to four of the domains (**Table 24**). To meet minimum alignment cutoffs by the typically accepted decision rules (used in this study), each reporting category would need to have at least six items in order to have some reliability in judging a student's proficiency for that reporting category. If each of the 10 domains are considered a reporting category, then another 33 items for the SAT Form April 2017 and 35 items for the SAT Form October 2017 would be needed to supplement the existing items to attain full alignment with the GSE for Algebra I. The required number of revised or replaced items (to attain minimum levels of alignment) could be lowered if the domains were grouped into larger categories, resulting in fewer reporting categories overall.

Both of the SAT forms had four or more items that mapped to four of the 10 domains—A.CED, A.REI, F.IF, and S.ID. In addition, SAT Form April 2017 had four items that mapped to Domain N.Q. The SAT Form April 2017 had two domains (A.CED and S.ID) and the SAT Form October 2017 had one domain (A.REI) with more than six corresponding items, a sufficient number to have an acceptable Categorical Concurrence. Nearly all of the 10 domains and the two SAT forms had an acceptable Depth-of-Knowledge Consistency. Only one domain on each form, A.BF on Form April 2017 and A.SSE on Form October 2017, did not. Thus, the items on the SAT forms were comparable in level of content complexity as expected by the GSE for Algebra I.

Both SAT forms had an acceptable Range-of-Knowledge Correspondence with two of the GSE for Algebra I domains, A.APR and A.CED. An acceptable range is attained for a domain if at least half of the underlying standards have one or more corresponding items. The Domain A.APR only had one underlying standard so range is less meaningful for that domain. The SAT Form April 2017 also had an acceptable range for Domains F.BF and S.ID. For about half of the domains, the number of domains with an unmet Range-of-Knowledge is related to some degree to the relatively low number of items corresponding to each domain. To a lesser degree, the unmet Range-of-Knowledge reflects the distribution of items among the standards underlying a domain. For example, the majority of reviewers found nearly three items on the SAT Form October 2017 that

mapped to standards under F.BF, but all three of these items corresponded to only one of the four underlying standards (MGSE9-12.F.BF.3). So for Domain F.BF, the Range-of-Knowledge Correspondence was not met. Thus, range between the SAT forms and the GSE for Algebra I was low for the majority of the 10 domains.

The low number of items on the SAT forms corresponding to each of the domains lessens the meaning of the Balance of Representation index values. All of the Balance indices were 0.70 or higher on both forms. For domains with multiple corresponding items, the balance index indicates that items did not over emphasize one standard more than others. However, the balance index is minimally informative considering the low numbers of items mapped to each domain.

Overall, the two SAT forms would need to be supplemented by 33-35 items to be considered fully aligned with the GSE for Algebra I. Generally, this number of items would be needed to have at least six items for each of the 10 domains. If there were fewer domains by aggregating the existing ones into larger clusters, then fewer items would be needed to include in a supplement assessment. The items on the two SAT forms, in general, had the same or higher DOK of the corresponding standards and compared favorably in complexity with the standards. Supplementary items would be needed to have adequate coverage of the content under the majority of the domains. With only half of their items mapping to Algebra I standards, the two SAT forms cannot not be considered to be aligned with the GSE for Algebra I.

Table 23. SAT Form April 2017 – mathematics with GSE for Algebra I

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
N.RN The Real Number System	0	--	--	--	NO	N/A	N/A	N/A
N.Q Quantities	4	81%	33%	--	NO	YES	NO	N/A
A.SSE Seeing Structure in Expressions	1	--	--	--	NO	N/A	N/A	N/A
A.APR Arithmetic with Polynomials & Rational Expressions	0	--	--	--	NO	N/A	N/A	N/A
A.CED Creating Equations	9	85%	93%	0.70	YES	YES	YES	YES
A.REI Reasoning with Equations & Inequalities	5	85%	30%	0.74	NO	YES	NO	YES
F.IF Interpreting Functions	5	70%	28%	0.88	NO	YES	NO	YES
F.BF Building Functions	2	--	--	--	NO	N/A	N/A	N/A
F.LE Linear, Quadratic, and Exponential Models	1	--	--	--	NO	N/A	N/A	N/A
S.ID Interpreting Categorical & Quantitative Data	7	78%	51%	0.83	YES	YES	YES	YES

*Number of items

Table 24. SAT Form October 2017 – mathematics with GSE for Algebra I

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
N.RN The Real Number System	0	--	--	--	NO	N/A	N/A	N/A
N.Q Quantities	1	--	--	--	NO	N/A	N/A	N/A
A.SSE Seeing Structure in Expressions	0	--	--	--	NO	N/A	N/A	N/A
A.APR Arithmetic with Polynomials & Rational Expressions	2	--	--	--	NO	N/A	N/A	N/A
A.CED Creating Equations	5	88%	50%	0.89	NO	YES	YES	YES
A.REI Reasoning with Equations & Inequalities	9	95%	29%	0.85	YES	YES	NO	YES
F.IF Interpreting Functions	4	65%	23%	0.85	NO	YES	NO	YES
F.BF Building Functions	2	--	--	--	NO	N/A	N/A	N/A
F.LE Linear, Quadratic, and Exponential Models	2	--	--	--	NO	N/A	N/A	N/A
S.ID Interpreting Categorical & Quantitative Data	4	53%	34%	0.92	NO	YES	NO	YES

*Number of items

Source of Challenge Issues and Reviewers' Comments: ACT and SAT Test Forms
Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have about an item. A Source-of-Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items for which more than one reviewer left a Source-of-Challenge comment.

Reviewers also wrote notes about many items on each form. For any item that did not match an Algebra I standard, reviewers made note of the general topic targeted by the item. Some notes indicate when only part of a particular standard was targeted by an assessment item. These notes also include general comments as well as indicate concerns with items. Some notes include suggestions for resolutions to editorial issues identified or suggestions for otherwise strengthening an item. After coding each assessment form, reviewers were asked to respond to four debriefing questions. The full text of reviewers' notes and debriefing comments can be found in **Appendices C** and **D**.

Reviewers left very few debriefing comments about the ACT or the SAT test forms, primarily noting that the test forms had limited overlap with the GSE for Algebra I.

Reliability among Reviewers
Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 25**, on the following page, were computed after adjudication. The overall intraclass correlation among the Algebra I reviewers' assignment of DOK levels to items was reasonably high for three of the four test forms analyzed (**Table 25**). The intraclass correlation for ACT Form 74C intraclass correlation (0.65) is lower than the value of 0.8 that generally indicates a high level of agreement among the reviewers. This lower agreement may relate to adjustments in coding process described below.

After reviewers began coding the first assessment (half were coding the ACT Form 74C and half were coding the SAT Form April 2017) the procedure for assigning DOK levels to the items that mapped to content other than Algebra I was modified. Initially, reviewers were instructed to assign a DOK level to all items on the assessment. Because this was slowing reviewers down and the information would not be used, the directors of the study changed the procedure by having the reviewers assign a 4 as the DOK level of items mapped to content other than the Algebra I standards. The WATv2 requires a reviewer to assign some DOK to each item so reviewers could not leave the DOK level blank. The directors decided to have reviewers assign a DOK level 4 to each item that was judged not to match any of the Algebra I standards. Because a DOK 4 would not apply to an assessment that is completed in a single sitting, the DOK 4 coding served as an identifier for non-Algebra I items. Some reviewers had completed coding ACT Form 74C or the SAT Form April 2017 before they received these new instructions. The statistic was calculated with some people assigning 4 to an item while others assigned a 1 or 2 which served to lower the intraclass correlation values. More reviewers in coding the SAT Form April 2017 were informed about the change and assigned a non-Algebra I item a DOK 4. The intraclass correlation values for ACT Form A10 (0.81) and SAT Form October 2017 (0.95) are more reflective of the agreement among reviewers. Even with

the change in procedure, the intraclass correlations for three of the four forms analyzed were relatively high.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other seven reviewers. For example, for eight reviewers a total of 28 comparisons were computed for each item. For most alignment studies, the standards pairwise agreement is higher than 0.6. The pairwise agreement for assigning standards to items was greater than 0.6 for all test forms analyzed. For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For the first two forms analyzed, ACT Form 74C and SAT Form April 2017, the pairwise domain agreement met the desired level. For the other two assessments, the reporting category was reasonably high for the other two assessment forms, both near 0.80 agreement. Reviewers will vary in their codings the more they have difficulty in finding a precise match between an assessment item and the standards.

Table 25. Intraclass and Pairwise Comparisons, ACT, and SAT with GSE Algebra I

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
ACT Form 74C	0.65	0.44	0.89	0.91
ACT Form A10	0.81	0.67	0.74	0.79
SAT April 2017	0.83	0.58	0.87	0.92
SAT Oct 2017	0.95	0.69	0.78	0.83

Summary of Comparisons of the Two Assessments

A summary of alignment results by test form is provided in **Table 26**. All test forms were found to need major adjustments in order meet minimum cutoffs for alignment with the GSE for Algebra I.

Table 26. Percent of GSE Reporting Categories for Algebra I with Acceptable Level on Each Alignment Criterion when Compared to Four Test Forms

Assessment Form	Categorical Concurrence (Percent of RCs with over six items)	Depth-of-Knowledge Consistency (50% at/above)	Range-of-Knowledge (50% of standards)	Balance of Representation (without possible weakness)
ACT 74C	0%	20%	0%	20%
ACT A10	0%	20%	10%	20%
SAT Apr 2017	20%	50%	20%	40%
SAT Oct 2017	10%	40%	10%	40%

Findings: Geometry

Framework Analysis for Mathematics – Geometry

Raven McCrory, a mathematics educator at Michigan State University, conducted a review of the design documents and other explanatory materials found for each of the ACT and SAT mathematics assessments. The mathematics framework analysis is included as **Appendix E** for Algebra I. Information from this report was used to compare the GSE for Algebra I and Geometry to content expectations as specified in design or interpretation documents from the ACT and the College Board. This analysis also provided information on the structure of the ACT and SAT and administration instructions for both. The design and interpretation documents included test blueprints, test specifications, and curriculum standards as were available.

Framework documents for both the SAT and ACT had standards that corresponded to fewer than 50% of the GSE for Geometry. The ACT CCRS had four standards that matched to the GSE for Geometry (9%) at the level of the benchmark for college and career readiness. The SAT Insight Skills had one standard that matched to the GSE for Geometry (2%) at the level of the benchmark for college and career readiness. Framework documents for both of the assessments did have standards related to levels higher than the benchmark that corresponded to about the same number for each assessment with the GSE for Geometry, 18 standards for ACT CCRS and 17 standards for SAT Insight Skills and higher level standards. Thus, about 40 or 49 percent of the GSE for Geometry had related standards to content specified in the framework documents for both assessments. Neither of the assessments' documents specified having students prove theorems, do constructions, or apply transformations, all of which are included in the GSE for Geometry. The assessments' documents also varied in how the expectations were described. The SAT documents frequently indicated that the geometry constructs should be used in solving problems whereas the ACT CCRS indicated that the students should be able to apply the properties and concepts without stating the context. For example, MGSE9-12.G.SRT.5 stated, "Use congruence and similarity criteria for triangles to solve problems…." The SAT Higher Level Standards that was judged to relate to this standard stated, "Use concepts and theorems about congruence and similarity to solve problems about lines, angles, and triangles…." The ACT CCRC Standard G603 that was judged to correspond to the Georgia standard stated, "Apply properties of 30°-60°-90°, 45°-45°-90°, similar, and congruent triangles." Other observed differences were in the detail that certain geometric ideas were described. The ACT CCRS was much more explicit in how it described expectations for Expressing Geometric Properties with Equations whereas the SAT documents were more detailed in stating expectations about constructs with circles. In general, the design and interpretation documents for the two assessments only corresponded to less than half of the GSE for Geometry and varied in the degree of detail that documents for each assessment described related expectations.

A comparison of session times, item counts, and item types are provided in **Table 27**. The mathematics assessments differed in their structure and the type(s) of items. The ACT mathematics assessment consisted of 60 items completed in 60 minutes, all items equally weighted at one point. All 60 items were multiple choice with five choices. Calculators were permitted for use when taking the ACT mathematics test but not

required. Students could use most calculators, including four-function, scientific, or graphing calculators except for those explicitly prohibited such as those with built-in or downloaded algebra computer system functionality.

The SAT mathematics assessment had 58 items administered in two parts, including 20 items where calculators were not permitted and 38 items where students were permitted to use a calculator. All items were equally weighted at one point. The College Board SAT website provides a list of brands and models of calculators that are acceptable for use on the mathematics test. Permitted calculators include most graphing calculators and all scientific calculators. More basic four-function calculators are permitted but not recommended. Students were allotted 80 minutes to complete the mathematics proportion of the assessment. The SAT assessments had two types of items, multiple choice (78%) and grid-ins (22%), in which students fill in a grid to enter a positive whole number, decimal, or fraction (**Table 27**). The Georgia Geometry EOC has 73 total items including a variety of item types. Of these, 52 items (for a total of 58 points) contribute to a student's final score, and the selected response items are all equally weighted at one point.

Table 27. Georgia EOC, ACT, & SAT Item Types–Mathematics

	Item Type										
	Multiple-choice		**Constructed Response (CR)**		**Extended CR**		**Technology-enhanced**		**Fill-in-the-grid**		**Total Number**
Test	N	%	N	%	N	%	N	%	N	%	
Geometry EOC	69	95	2	3	1	1	1	1	--	--	73
ACT	60	100	--	--	--	--	--	--	--	--	60
SAT	45	78	--	--	--	--	--	--	13	22	58

Table 28. Time per assessment item/task for Georgia EOC, ACT, and SAT–Mathematics

Test	**Number of Items**	**Assessment Time**	**Average Time per Item****
Geometry EOC	73 items*	170 min	2.3 min
ACT Form A10	60 items	60 min	1 min
SAT Apr 2017	58 items	80 min	1.4 min

*Of these, 11 items are field test items and only some of the Norm-Referenced Test items (those that correspond to GSE) will contribute to a student's Criterion-Referenced score; total item count was included to calculate average time per item.
**Note that some items may take more time than others.

Standards

The 45 Georgia Geometry standards were clustered under seven domains (**Table 29**). Unlike the Algebra I analysis, no elements (parts of standards) were included in the Geometry analysis. Six of the domains provided expectations on geometric concepts and ideas and one domain expressed knowledge related to probability. Reviewers were asked to review assigned DOKs to standards used in previous studies. They were asked to change any DOK they disagreed with and to assign a DOK to any element that did not have a DOK. A consensus process was used. The reviewers did not change any of the DOK values assigned to the Geometry standards. They added the DOK for two standards, MGSE9-12.G.C.4 and 5. Standard G.C.4 was assigned a DOK 2 and Standard G.C.5 was assigned a DOK 3. Nearly 60% of the Georgia Geometry standards were judged to require students to apply skills and demonstrate conceptual knowledge (DOK 2). Another 30% of the standards expected students to engage in strategic thinking and complex reasoning (DOK 3). One standard, MGSE9-12.G.MG.3 (apply geometric methods to solve design problems), was judged to require extended thinking (DOK 4) and four standards were judged to be recall of information (9%).

Table 29. Percent of Expectations by Depth-of-Knowledge (DOK) Levels for the Mathematics Georgia Standards of Excellence for Geometry

Domain	Total Number of Standards	DOK Level	Number of Standards by Level	Percent within Conceptual Category by Level
G.CO Congruence	13	1 2 3	1 8 4	7.69 61.54 30.77
G.SRT Similarity, Right Triangles, and Trigonometry	8	2 3	5 3	62.5 37.5
G.C Circles	5	1 2 3	1 2 2	20 40 40
G.GPE Expressing Geometry Properties with Equations	5	1 2 3	1 2 2	20 40 40
G.GMD Geometric Measurement and Dimension	4	2 3	2 2	50 50
G.MG Modeling with Geometry	3	2 4	2 1	66.67 33.33
S.CP Conditional Probability and the Rules of Probability	7	1 2 3	1 5 1	14.29 71.43 14.29
Total	45	1 2 3 4	4 26 14 1	9 58 31 2

Mapping of Items by Standards

If no particular grade-level standard is targeted by a given assessment item, reviewers were instructed to code the item at the domain level. This coding to a generic standard generally indicated that the assessment item did not precisely target one of the standards included in the study and may be above or below grade level. However, if the item is grade-appropriate, then this situation may instead indicate that there is a part of the content not expressly or precisely described in the standards, or that there is a part of the content within the standards that is being interpreted differently by different parties. Items coded to generic standards may highlight areas in the standards with missing content or where the statement of the standard is not as precise as it should be as well as a mismatch with an assessment.

A majority of the reviewers coded seven items on ACT Form 74C and six items on ACT Form A10 to a generic standard. Most of items on ACT Form 74C assigned to a generic standard were judged to correspond to two domains, G.GMD (Geometric Measurement and Dimension) and S.CP (Conditional Probability and Rules of Probability). Reviewers' main reason why these standards did not fit the high school geometry standards was because they corresponded more closely with middle school standards (e.g. an item requiring the area for a rectangle was judged to relate more to middle school standards rather than the high school geometry standard). Similarly for ACT Form A10, the main reason reviewers gave for assigning items to a generic standard was because these items corresponded more to middle school or below standards.

The majority of the reviewers only found three items on the SAT forms, two on the April 2017 Form and one on the October 2017 Form, that did not fit with the high school geometry standards. These three items more closely related to middle school standards. Two required knowledge of a circumference of a circle and one knowledge of the volume of a rectangular prism. Reviewers found a higher proportion of items on the two ACT forms that did not precisely correspond to the high school geometry standards than on the two SAT forms.

Table 30. Items Assigned to Generic Content Expectations by a Majority of Reviewers for the Mathematics GSE for Geometry Alignment Analysis

Test	Generic Content Expectation	Item Number (N Reviewers)	Comments
ACT Form 74C	G.GPE	30(5), 31(6)	[Information subject to nondisclosure agreements has been omitted for public release.]
	G.GMD	1(8), 18(8), 43(6)	[Information subject to nondisclosure agreements has been omitted for public release.]
	S.CP	2(8), 26(8)	[Information subject to nondisclosure agreements has been omitted for public release.]
ACT Form A10	G.CO	35(7)	[Information subject to nondisclosure agreements has been omitted for public release.]
	G.GMD	8(7), 23(6), 59(5)	[Information subject to nondisclosure agreements has been omitted for public release.]
	S.CP	2(7), 42(7)	[Information subject to nondisclosure agreements has been omitted for public release.]
SAT April 2017	G.C	57(5), 58(5)	[Information subject to nondisclosure agreements has been omitted for public release.]
SAT October 2017	G.GMD	16(6)	[Information subject to nondisclosure agreements has been omitted for public release.]

Neither assessment had items that targeted a high percentage of the GSE for Geometry. At most, ACT Form 74C had items that corresponded to about one-third of the Georgia Geometry Standards (**Table 31**). On the other three assessment forms, reviewers only found items that corresponded to about one-fifth of the Geometry GSE. All four assessments had items that reviewers judged to map to generic standards, generally because these items corresponded best to standards from lower grades.

The two ACT assessment forms had a higher percentage of items, 35% and 28% (about one third of the items), that reviewers judged corresponded to geometry standards (**Table 32**). The two SAT assessment forms only had eight or nine items (about 15% of the items) that reviewers found corresponded to geometry standards. Even though ACT Form A10 had more items judged to target geometry standards, the number of GSE for Geometry targeted was about the same as for the two SAT forms. Consequently, the findings for the number of standards with corresponding items with geometry standards was mixed for the two ACT forms and fairly consistent between the two SAT forms. The general conclusion is that both assessments only addressed in some way less than a third of the GSE for Geometry. The two ACT forms had nearly twice the number of items than the two SAT forms that were related to geometry, but this did not necessarily mean that more geometry content (as described by the GSE for Geometry) was assessed by the ACT forms.

Table 31. Number and Percent of Mathematics GSE for Geometry with at least One Corresponding Item Found by a Majority of Reviewers

Test	Total Number of Items	Number of GSE Targeted	Number of Geometry Standards with Corresponding Item(s)
ACT Form 74C	60	15*	31%
ACT Form A10	60	11*	23%
SAT Apr 2017	58	8**	17%
SAT October 2017	58	10**	22%

* Including three generic standards
** Including one generic standard

Table 32. Number and Percent of Mathematics Items for ACT and SAT Assessments Judged by Majority of Reviewers as Corresponding to Mathematics GSE for Geometry Standards

Test	Total Items	Items Corresponding to Geometry Standards		Indecisive		Items Corresponding to Other Mathematics Content	
		Number	Percent	Number	Percent	Number	Percent
ACT Form 74C	60	21	35%	1	2%	38	63%
ACT Form A10	60	17	28%	2	3%	41	68%
SAT Apr 2017	58	8	14%	1	2%	49	84%
SAT Oct 2017	58	9	16%	--	--	49	84%

Although both assessments had nearly the same number of items (58 or 60 items; **Table 32)**, the assessments varied in the proportion of items that targeted Geometry standards. Across both ACT test forms, a majority of reviewers found that around 30% of the total items mapped to at least one of the GSE Geometry standards. Across both SAT test forms, a majority of reviewers found a lesser proportion of items, 15%, that mapped to at least one GSE Geometry standard.

Comparison of Overall DOK Distribution

A comparison of the overall DOK distribution for items on each assessment, averaged across the two test forms, is shown in **Table 33**. The ACT and SAT assessments analyzed were very similar in the content complexity of the geometry items. Most of the geometry items (over three-quarters) were judged to have a DOK 2, related to conceptual understanding and engaging with mathematical relationships. Only one or two of the items were judged to require students to use strategic thinking (DOK 3). Eighteen percent of the geometry items on both assessments were judged to be recall of information (DOK 1).

Table 33. DOK Distribution of Geometry Items, averaged across two test forms, for the ACT and SAT

Test	DOK 1	DOK 2	DOK 3
ACT	18%	79%	3%
SAT	18%	76%	6%

Alignment Statistics and Findings for ACT and SAT Test Forms and GSE for Geometry

Overall alignment results are summarized in **Table 34** below and then detailed for each test form in the pages that follow. A main alignment issue for all forms was a lack of Categorical Concurrence for any of the reporting categories. Both ACT and both SAT test forms were found to need major adjustments to meet minimum cutoffs for alignment.

Table 34. Overall Alignment Findings for Two Forms Each of ACT and SAT Mathematics Assessments with GSE for Geometry

Test Form	Alignment Findings	Number of Items that Need Revision/Replacement for Minimum Alignment
ACT 74C	Needs Major Adjustments	23-30*
ACT A10	Needs Major Adjustments	29
SAT April 2017	Needs Major Adjustments	34
SAT Oct 2017	Needs Major Adjustments	33

*30 items to account for replacement of below-grade items

Results by Test Form
The results of the analysis for each of the four alignment criteria are provided in **Tables 35** to **38** for each mathematics test form for the seven Geometry Reporting Categories. The approximate numbers of replaced or revised items necessary to meet minimum levels of alignment are provided for each test form. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form. The reviewers' notes and debriefing comments (**Appendices C** and **D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In **Tables 35** to **38**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth-of-Knowledge Consistency, 10% under an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation. Categorical Concurrence is reported in average number of items. Depth of Knowledge Consistency is reported by the percent of items that were at or above the DOK of the corresponding standard. Range-of-Knowledge is reported as the percent of standards within each reporting category that were targeted by one or more items. Balance of representation is an index value, ranging from 0-1. Alignment statistics for DOK Consistency, Range-of-Knowledge, and Balance are reported only for reporting categories that have three or more corresponding items.

ACT Test Forms
The two ACT forms analyzed varied some in the number of items that were judged to correspond to GSE for Geometry. The majority of reviewers found 21 items on Form 74C that mapped to 15 geometry standards, including three generic standards. They found 17 items on Form A10 that mapped to 11 geometry standards, including three generic standards. Thus, one-third or fewer of the GSE for Geometry were addressed in any way by each of the two ACT test forms. Only two of the GSE for Geometry had at least one corresponding item on both of the ACT forms, MGSE9-12.G.CO.9 and MGSE9-12.G.CO.10 (proving theorems about lines, angles and triangles), and one generic standard, MGSE9-12.S.CP. This suggests that over a number of ACT forms a range of Georgia geometry standards could be targeted. For the two ACT forms analyzed nearly 40 percent of the standards were targeted, including three generic standards.

ACT Form 74C
Six of the items on ACT Form 74C were double or even triple coded by more than one reviewer. Multiple codings were used generally when a reviewer did not find an exact fit between what knowledge students had to use to answer the item correctly and a standard. The items that were multiple coded, rather than being more robust items measuring more than one idea, were items that did not fit as well the GSE for Geometry. The average total number of hits of the 60 items on ACT Form 74C was 65.26 (Appendix B), 27.51 coded by the eight reviewers to Geometry standards and 37.75 coded to mathematical topics other than Geometry. Of the 38 items judged by the majority of the reviewers to assess other topics, 15 items were noted as related to Algebra I, 11 items as related to number, nine items as related to more advanced mathematics such as

trigonometry, two items as related to data and statistics, and one item required students to come up with a logical argument. Item 48 that was related to classifying triangles was coded by half of the reviewers as a geometry item and the other half as another mathematics topic item because it targeted a middle school standard.

With each of the seven domains for geometry considered as a reporting category, the alignment between ACT Form 74C and the GSE for Geometry would need major improvement to be considered acceptable. Six or more items per reporting category is the number typically considered sufficient to make a reliable decision about a student's proficiency. This cutoff of six items was met for two (G.CO [congruence] and S.CP [conditional probability]) of the five domains (Table xx). The ACT Form 74C nearly had a sufficient number of items to have an acceptable level on the Categorical Concurrence criterion for one more domain (G.GMD [measurement and dimensions). It should be noted that two items corresponding to S.CP and two corresponding to G.GMD were mapped to generic standards. Reviewers indicated these items better fit with middle school standards. For the other four domains, the average number of items coded to a domain was 3.25 for two (G.SRT, similarity, right triangle, and trigonometry, and G.GPE, expressing geometric properties with equations) or not tested (fewer than two corresponding items) for another two domains (G.C, circles, and G.MG, modeling with geometry).

The level of complexity of the items on the ACT Form 74C was generally weakly acceptable. Only the items found corresponding to the domain S.CP had a sufficient proportion of items with a DOK level that was the same or higher than the DOK level of the corresponding standard, 50 percent or higher. For the other six domains, three domains had between 40 and 50 percent of the items with DOK levels that were at or above the level of the corresponding standard which is considered weakly acceptable for DOK consistency. One domain only had corresponding items with a DOK level below the level of the corresponding standard, and two domains were not considered tested.

Range-of-Knowledge Correspondence criterion was not met for any of the seven domains. To be acceptable the assessment had to have at least one item that corresponded to at least half of the standards under a domain. For the domains that were considered tested, range was weakly acceptable for four domains and not acceptable for one domain. The items that did correspond to each of the domains were adequately distributed among the standards without any one standard being considered over emphasized to have an acceptable Balance of Representation.

Overall, using the acceptable levels for each criterion as discussed, for ACT Form 74C and the GSE for Geometry to have full alignment would require supplementing the existing items with 23 additional items. Most of these items (N=18) would be needed in order for the assessment to have at least six items for each of the seven domains. If the items were carefully selected to have a level of complexity matching that of the corresponding standard and targeting standards not currently assessed, then most of the concerns for DOK Consistency and Range-of-Knowledge Correspondence would be resolved. The majority of the reviewers found seven of the 21 items that mapped in some way to the GSE for Geometry to be below grade level. For the assessment to be on grade level, then these seven items should be replaced or another seven items added to supplement the existing assessment. Thus, a supplementary assessment of

about 30 items would be needed for ACT Form 74C and the GSE for Geometry to have acceptable alignment. The number of items for a supplementary assessment could be reduced if the number of reporting categories were reduced to two or three.

In their debriefing notes, reviewers specifically identified general topics of similarity, congruence, proof, and modeling as areas that were in the standards but not significantly assessment by the ACT Form 74C. They had some difficulty finding a close match of items on the assessment with the GSE for Geometry. One reviewer summarized, "…it was very difficult to map an item onto a standard for one or more of the following reasons: inappropriate content (e.g. algebra standard [not geometry]); inappropriate grade level (below high school grade level); the item covered only entry level information needed to perform the standard; the item contained a single word mentioned in the standard but did not focus on the intent of the standard; or the distractors reduced the complexity of the task…."

Table 35. ACT Form 74C June 2017 – mathematics with GSE for Geometry

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
G.CO Congruence	6	41%	39%	0.84	YES	NO	NO	YES
G.SRT Similarity, Right Triangle, and Trigonometry	3	43%	41%	0.83	NO	NO	NO	YES
G.C Circles	1	--	--	--	NO	N/A	N/A	N/A
G.GPE Expressing Geometry Properties with Equations	3	45%	43%	0.87	NO	NO	NO	YES
G.GMD Geometric Measurement and Dimension	5	0%	40%	0.71	NO	N/A	N/A	YES
G.MG Modeling with Geometry	0	--	--	--	NO	N/A	N/A	N/A
S.CP Conditional Probability and the Rules of Probability	7	86%	47%	0.81	NO	YES	NO	YES

ACT Form A10

Although there is some variation in the alignment results between the two ACT forms analyzed, the general findings are very similar. The alignment between the ACT Form A10 and the GSE for Geometry would need major improvement or to be significantly augmented to be considered aligned based on the decision rules used for this study. Reviewers did not find on the ACT Form A10 six items that corresponded to any of the seven geometry domains or reporting categories (**Table 36**). Reviewers did find five items that corresponded to standards under each of two domains (G.CO and C.GMD). However, three of the items that mapped to Domain C.GMD were considered below grade level. Reviewers also found Domains G.C and G.MG not to be tested and Domains G.SRT and G.GPE to be moderately assessed with three items. This distribution of items on Form A10 among the domains was very similar to the distribution of items on Form 74C. The two forms did vary in the assessment of standards under Domain S.CP (conditional probability). Reviewers only found nearly four items on Form A10 compared to nearly eight items on Form 74C that mapped to standards under this domain.

The DOK consistency was acceptable for only two of the seven domains and the ACT Form A10. Just over half of the three items that mapped to standards under Domain G.SRT and nearly all of the four items that mapped to standards under Domain S.CP had a DOK level that was the same or greater than the DOK level of the corresponding standard. The other five domains were not tested, had a weak DOK consistency, or had fewer than 40% of the items with the same or higher DOK level than the level of the corresponding standard.

ACT Form A10 and the seven domains of the GSE for Geometry did not have an acceptable level for the Range-of-Knowledge Correspondence for any of the domains. Range was weakly met for the assessment and G.GPE with two of five standards with corresponding items, but for the other six domains, either the domain was not tested or only about one-fourth of the underlying standards with at least one corresponding standard. The range criterion was more of an issue for Form A10 than for Form 74C. Whereas range was weakly met for four domains with Form 74C, range was only weakly acceptable for one domain and not met for four other domains. In general, the items that corresponded to domains were distributed among the standards under the domain to have an acceptable balance.

Overall, using the acceptable levels for each criterion as discussed, for ACT Form A10 and the GSE for Geometry to have full alignment would require supplementing the existing items with approximately 29 additional items. As with Form 74C most of these items (at least 23 items) would be needed in order for the assessment to have at least six items for each of the seven domains. The supplementary items would need to be selected taking into consideration standards not assessed and the level of complexity.

In their debriefing comments, reviewers noted major topics that were not addressed by the assessment including transformation of geometric figures, similarity, congruence, arc and angle measurement, conditional probability, volume of composite figures, and

reasoning using theorems. As for the other ACT form, the alignment between the GSE for Geometry and the ACT Form A10 would need major improvement or to be augmented by an assessment with nearly half of the number of items as on the original assessment to be considered as acceptably aligned.

Table 36. ACT Form A10 June 2017 – mathematics with GSE for Geometry

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
G.CO Congruence	5	43%	27%	0.84	NO	NO	NO	YES
G.SRT Similarity, Right Triangle, and Trigonometry	3	53%	26%	0.94	NO	YES	NO	YES
G.C Circles	0	--	--	--	NO	N/A	N/A	N/A
G.GPE Expressing Geometry Properties with Equations	3	34%	45%	0.78	NO	NO	NO	YES
G.GMD Geometric Measurement and Dimension	5	2%	22%	0.96	NO	NO	NO	YES
G.MG Modeling with Geometry	0	--	--	--	NO	N/A	N/A	N/A
S.CP Conditional Probability and the Rules of Probability	3	92%	27%	0.81	NO	YES	NO	YES

SAT Test Forms

On the two SAT forms analyzed, reviewers only found eight or nine items (about 15 percent of the items on each form) that mapped to standards under the GSE for Geometry (**Tables 37-38**). These items were found to relate to about one-fifth of the 45 GSE for Geometry. On each of the two forms, the majority of the reviewers coded one or two items to generic standards because these items targeted content knowledge that was below high school level. The few geometry items on each of the SAT forms mapped to seven of the same standards. This suggests that the two forms had some parallel construction in content covered as related to geometry.

SAT Form April 2017

The SAT Form April 2017 and the GSE for Geometry had very minimal alignment. Only eight items on the form targeted at least one of the 45 geometry standards. Reviewers did find four items that the majority coded to more than one standard—Items 5, 48, 49, and 50. However, SAT Form April 2017 would need to be supplemented with at least 34 additional items to be minimally aligned with the GSE for Geometry based on the decision rules for alignment used in this study and described in this report. SAT Form April 2017 did not have at least six items that corresponded to any of the seven domains or reporting categories of the GSE for Geometry. Nearly three items mapped to each of three domains—G.CO, G.SRT, and G.GPE. Of these items, two were mapped by the majority of reviewers to generic standards. Reviewers' comments indicated that these items required students to have knowledge of some middle school content, but then included a part that was at the high school level (e.g. exponential growth or creating an equation). The form had too few items corresponding to the other four domains for these domains to be considered tested.

None of the seven domains had an acceptable Range-of-Knowledge Correspondence. Only one-fifth or just over one-third of the standards under those domains tested had corresponding items. The low range was related to having a low number of geometry items on the assessment. As for Algebra I, the balance index is minimally informative considering the low numbers of items mapped to each domain.

Overall, the alignment between the SAT Form April 2017 and the GSE for Geometry would need major improvement to be aligned. The test form would need to be augmented with at least 34 items to reach full alignment with the seven domains as reporting categories. These items would have to be selected considering complexity and targeting underlying standards that are not currently assessed. The number of supplemental items could be reduced if the number of reporting categories were reduced to two or three.

In their debriefing comments, reviewers indicated that the SAT assessment tested more knowledge of algebra than geometry. There were a number of geometry topics included in the Georgia standards that were not addressed on the SAT Form April 2017 according to one reviewer: transformations of geometric figures in the coordinate plane, similar and congruent triangles, volume of three-dimensional figures, arc and angle measure relationships in circles, and geometric constructions. The reviewers found most of the geometry items to have a DOK level 2, which was similar to the GSE for Geometry. The reviewers noted that the SAT Form April 2017 had fewer geometry items than the ACT, although both the SAT and the ACT assessments did not fit very well with the GSE for Geometry.

Table 37. SAT Form April 2017 – mathematics with GSE for Geometry

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
G.CO Congruence	3	19%	24%	0.98	NO	NO	NO	YES
G.SRT Similarity, Right Triangle, and Trigonometry	2	--	--	--	NO	N/A	N/A	N/A
G.C Circles	1	--	--	--	NO	N/A	N/A	N/A
G.GPE Expressing Geometry Properties with Equations	2	--	--	--	NO	N/A	N/A	N/A
G.GMD Geometric Measurement and Dimension	0	--	--	--	NO	N/A	N/A	N/A
G.MG Modeling with Geometry	0	--	--	--	NO	N/A	N/A	N/A
S.CP Conditional Probability and the Rules of Probability	1	--	--	--	NO	NO	NO	NO

SAT Form October 2017

The SAT Form October 2017 and GSE for Geometry had about the same degree of alignment as for the SAT Form April 2017 with variation for only about two items. As for the April 2017 Form, the alignment was very low. The majority of reviewers did find nine of the 58 items (16%) on SAT Form October 2017 that corresponded in some way with Georgia geometry standards. Because reviewers coded some of these items (Items 27 and 51) as mapping to more than one standard, these nine items corresponded to 10 standards. Reviewers did not find any of the seven domains that had six or more items from the SAT Form October 2017 that targeted underlying standards. At most reviewers found two or three items that mapped to standards under any of the domain. The other four domains were not considered assessed. Thus, none of the seven domains and the GSE for Geometry met the minimum cutoff for the Categorical Concurrence criterion. The other alignment criteria are minimally informative because of the very few items that mapped to geometry GSE.

Overall, the SAT Form October 2017 would need to be supplemented by an additional 33 items to attain minimum accepted levels of alignment. These items generally would need to have a DOK level 2 and target standards not targeted by the current items. In the debriefing comment, one reviewer identified topics that were either partially or not at

all addressed by items on the assessment, “Major topics that were only partially covered by the assessment items were theorems about lines, angles, and triangles; right triangle trigonometry; arc and angle measure relationships in circles; and conditional probability. Major topics that did not have any corresponding items were transformations of geometric figures in the coordinate plane; theorems about parallelograms; geometric constructions; triangle similarity and congruence; arc length and sector area in circles; equations of circles; volume of geometric solids that are not rectangular prisms; geometric modeling; addition rule in probability; subsets of a sample space.” Most of the other reviewers identified a proportion of topics from this list. Reviewers felt that the alignment between the SAT Form October 2017 and the GSE for Geometry was similar with the other SAT form analyzed; both were very minimally aligned. One reviewer summarized, “It is difficult to see how this test could serve as a proxy for an EOC assessment in geometry.”

Table 38. SAT Form October 2017 – mathematics with GSE for Geometry

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
G.CO Congruence	1	--	--	--	NO	N/A	N/A	N/A
G.SRT Similarity, Right Triangle, and Trigonometry	2	--	--	--	NO	N/A	N/A	N/A
G.C Circles	1	--	--	--	NO	N/A	N/A	N/A
G.GPE Expressing Geometry Properties with Equations	2	--	--	--	NO	N/A	N/A	N/A
G.GMD Geometric Measurement and Dimension	1	--	--	--	NO	N/A	N/A	N/A
G.MG Modeling with Geometry	0	--	--	--	NO	N/A	N/A	N/A
S.CP Conditional Probability and the Rules of Probability	2	--	--	--	NO	NO	NO	NO

Source of Challenge Issues and Reviewers' Comments: ACT and SAT Test Forms

Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have about an item. A Source-of-Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items for which more than one reviewer left a Source-of-Challenge comment.

Reviewers also wrote notes about many items on each form. For any item that did not match a Geometry standard, reviewers made note of the general topic targeted by the item. Some notes indicate when only part of a particular standard was targeted by an assessment task. These notes also include general comments as well as indicate concerns with items. Some notes include suggestions for resolutions to issues identified. After coding each assessment form, reviewers were asked to respond to four debriefing questions. The full text of reviewers' notes and debriefing comments can be found in **Appendices C** and **D**. Reviewers' debriefing comments note the minimal alignment between the test forms and the GSE for Geometry, also noting that the tests analyzed are not intended as Geometry tests but rather as college entrance tests.

Reliability among Reviewers

Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 39**, on the following page, were computed after adjudication. The overall intraclass correlation among the Geometry reviewers' assignment of DOK levels to items was high for all test forms analyzed (**Table 39**). The intraclass correlation for all test forms was higher than the value of 0.8 that generally indicates a high level of agreement among the reviewers. The lowest intraclass correlation (for SAT April 2017) is a little lower than the other three forms, but still reasonably high. Similar to the Algebra I study, the process for assigning DOK was adjusted after the study start, and this lowest correlation corresponds to the test form that was coded before the adjustment, including DOK assignments for non-geometry items.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other seven reviewers. For example, for eight reviewers a total of 28 comparisons were computed for each item. For most alignment studies, the standards pairwise agreement is higher than 0.6. The pairwise agreement for assigning standards to items was much greater than 0.6 for all test forms analyzed. For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. This level was attained for all forms except the ACT Form A10, which was just under that agreement (0.88).

Table 39. Intraclass and Pairwise Comparisons, ACT, and SAT with GSE Geometry

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
ACT Form 74C	0.90	0.70	0.82	0.93
ACT Form A10	0.96	0.80	0.75	0.88
SAT Apr 2017	0.86	0.64	0.88	0.94
SAT October 2017	0.98	0.93	0.92	0.97

Summary of Comparisons of the Two Assessments

A summary of alignment results by test form is provided in **Table 40**. All test forms were found to need major adjustments in order meet minimum cutoffs for alignment with the GSE for Geometry. The main alignment issues for both the ACT and SAT with the GSE for Geometry is that the assessments have too few items to produce information that can be used to reliably make judgements about students' proficiency related to the Geometry Standards and the items on the assessment do not address enough breadth in the content as expected by the standards.

Table 40. Percent of GSE Reporting Categories for Geometry with Acceptable Level on Each Alignment Criterion when Compared to Four Test Forms

Assessment Form	Categorical Concurrence (Percent of RCs with over six items)	Depth-of-Knowledge Consistency (50% at/above)	Range-of-Knowledge (50% of standards)	Balance of Representation (without possible weakness)
ACT 74C	17%	17%	0%	71%
ACT A10	0%	34%	0%	71%
SAT Apr 2017	0%	0%	0%	17%
SAT Oct 2017	0%	0%	0%	0%

Findings: Biology

Framework Analysis for Science / Biology

Ms. Evans' framework analysis for science assessments revealed substantial differences between the ACT science test and the Georgia Biology EOC as pertains to test purpose, assessment targets, item type(s), allotted test time, and number of test items/tasks.

The framework analysis found that the ACT College and Career Readiness (CCR) Standards had very little match with the biology GSE. Partial matches were identified, focusing on the science and engineering practices components of the GSE. However, while the GSE contextualizes these practices within biology content, the ACT CCR Standards are contextualized across biology, chemistry, physics, and Earth/space science content.. Additionally, the ACT CCR Standards often overlapped with only a portion of the practices.

A comparison of session times, item counts and types, and science subjects included are provided in **Table 41**. While the ACT included multiple choice items only and includes biology, Earth/space, physics, and chemistry topics, the Georgia Biology EOC includes both multiple choice and technology enhanced items and targets biology only. The ACT also focuses centrally on student proficiency in science practice (contextualized within a variety of science content areas) while the Georgia Biology EOC is designed to measure student proficiency related to specific biology content in the context of science practices. Because of the differences in item counts and allotted test time, students would have twice as much time per item (1.8 minutes/item versus 0.9 min/item) on the Biology EOC compared with the ACT.

Table 41. Comparison of Georgia Biology EOC and ACT Session Times, Item Counts, Types, and Science Subjects Included

	Georgia Biology EOC	**ACT Science Test**
Assessment Time	2 sections; 70 min each TOTAL: 140 min maximum	1 section; 35 min TOTAL: 35 min maximum
Number of Items	76 items	40 items
Type of Items	Multiple Choice Technology Enhanced	Multiple Choice
Science Subjects Included	Biology	Biology, Earth/Space, Physics, Chemistry

Source: Georgia Department of Education, 2017; ACT, 2014

The full framework analysis can be found in **Appendix E** for Biology.

Standards

Study results are reported according to six reporting categories (RCs) corresponding to the six biology GSE. Each standard includes three to five objectives.

A summary of the levels of complexity within the GSE for Biology is given in **Table 42**. None of the standards included in the biology study (0%) were considered DOK 1. The majority of standards (67%) were considered a DOK level 2, emphasizing work that requires underlying conceptual understanding, as well as consideration of relationships between and among different ideas, connecting ideas, and more. Six standards (25%) were considered to be DOK 3, emphasizing expectations for abstract and hypothetical thinking, non-routine problem solving, deep analyses of data and other work. Three standards (13% percent) were considered DOK 4. A DOK 4 expectation is one that is both at least as complex as a DOK 3 but also requires extended time—days, weeks, or months—to complete. Although some components of these DOK 4 standards may be reasonably assessed by on-demand assessments, DOK 4 standards should not be expected to be fully assessed by an on-demand test.

Table 42. Expectations by Depth-of-Knowledge (DOK) Levels for GSE for Biology, February, 2018

Biology	Total Number of Expectations	DOK Level	Number of Standards by Level	Percent within RC by Level
GSE.SB1. Structures and Functions in Living Cells	5	2 3 4	3 1 1	60 20 20
GSE.SB2. Genetic Information Expressed in Cells	3	2 3	1 2	33.33 66.67
GSE.SB3. Biological Traits Passed on to Successive Generations	3	2	3	100
GSE.SB4. Interacting Systems within Single-Celled & Multi-Celled Organisms	3	2 3	2 1	66.67 33.33
GSE.SB5. Interdependence of All Organisms on One Another and Their Environment	5	2 3 4	2 2 1	40 40 20
GSE.SB6. Assess the Theory of Evolution	5	2	5	100
Total	24	2 3 4	16 6 2	67 25 8

Mapping of Items by Standards

Table 43 shows the items for each assessment that a majority of reviewers coded to a Biology GSE. This table also shows the number of items that a majority of reviewers coded to the category of "Science" –indicating that the item was related to the sciences but not related to any of the GSE for biology (e.g. was grounded in physical sciences). On the ACT, across all three forms analyzed, reviewers were unable to find a Biology GSE standard match for 31-39 items (~77-97% of total items). Reviewers were required to write an explanation in the case of assigning an item to a generic standard. These notes can be found in **Appendix C,** with any sensitive content subject to nondisclosure agreements omitted for public release.

Table 43. ACT Items Assigned to Biology GSE and to "Other Science", February, 2018

ACT Science Test/Form	Number of Items Correlated with Biology GSE (%)	Number of Items Correlated with "Other Science" (%)
ACT 74C	7 (17.5%)	33 (82.5%)
ACT A10	1 (2.5%)	39 (97.5%)
ACT 74H	9 (22.5%)	31 (77.5%)

Alignment Statistics and Findings for ACT Science Assessments Forms with GSE for Biology

Because only 2.5-17.5% of the items on the ACT science assessments correlated to the Biology GSE, the ACT science test cannot be considered aligned with the Biology GSE. This finding indicates only that the ACT Science test does not target the Biology GSE; it does not have any relevance to the quality of the ACT Science test or the alignment of the ACT Science test with the test's assessment targets (which are not related specifically to biology content). In other words, it is very reasonable that the ACT Science test does not align to any degree with the Biology GSE because the ACT Science test is not intended as a biology test.

Results by Test Form

The results of the analysis for each of the four alignment criteria are provided in **Tables 44** to **46.** for each ACT Science test form for Reporting Categories 1-6. The approximate numbers of replaced or revised items necessary to meet minimum levels of alignment are provided for each test form. More detailed data on each of the criteria are given in **Appendix B**, in the first three tables for each test form. The reviewers' notes and debriefing comments (**Appendices C** and **D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In **Tables 44** to **46**, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth-of-Knowledge Consistency, 10% under an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation.

ACT Form 74C

For ACT Form 74C, reviewers found items that corresponded to only one of the Georgia Standards of Excellence for Biology: **GSE.SB3.** *Obtain, evaluate, and communicate information to analyze how biological traits are passed on to successive generations.* All alignment criteria were met for this reporting category. Reviewers did not find any items on the test form that targeted any other standards. Consequently, for ACT Form 74C, a minimum of 30 items would need to be added or replaced to meet the minimum levels of acceptable alignment. These items would need to target the other five standards, with items that match the DOK of the standards and assess at least half of the elements within each standard.

Table 44. ACT Form 74C June 2017 – Science

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
GSE.SB1	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB2	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB3	6	100%	58%	0.70	YES	YES	YES	YES
GSE.SB4	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB5	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB6	0	--	--	--	NO	N/A	N/A	N/A

*Number of items

ACT Form A10

For ACT Form A10, reviewers found only one item that corresponded to the GSE for biology. Consequently, for ACT Form A10, a minimum of about 35 items would need to be added or replaced to meet the minimum levels of acceptable alignment.

Table 45. ACT Form A10 December 2017 – Science

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
GSE.SB1	1	--	--	--	NO	N/A	N/A	N/A
GSE.SB2	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB3	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB4	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB5	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB6	0	--	--	--	NO	N/A	N/A	N/A

*Number of items

ACT Form 74H

For ACT Form 74H, reviewers found items that corresponded to only one of the Georgia Standards of Excellence for Biology: **GSE.SB4.** Obtain, evaluate, and communicate information to illustrate the organization of interacting systems within single-celled and multi-celled organisms.

Although the minimum level of Categorical Concurrence was met for this reporting category, at least three of items would need to be revised or replaced to meet DOK consistency. Reviewers found only one item on the test form that targeted another Biology standard. Consequently, for ACT Form 74C, a minimum of 32 items would need to be added or replaced to meet the minimum levels of acceptable alignment.

Table 46. ACT Form 74C June 2017 – Science

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
GSE.SB1	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB2	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB3	0	--	--	--	NO	N/A	N/A	N/A
GSE.SB4	6	12%	42%	0.90	YES	NO	WEAK	YES
GSE.SB5	1	--	--	--	NO	N/A	N/A	N/A
GSE.SB6	0	--	--	--	NO	N/A	N/A	N/A

*Number of items

Source of Challenge Issues and Reviewers' Comments
Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have about an item. A Source-of-Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items on the ACT Science test forms for which a reviewer left a Source-of-Challenge comment.

Reviewers' notes specify the content area addressed in questions that they coded to the non-biology general science category. Reviewers noted items that had other science contexts (e.g. chemistry or physics). They also noted that some items assessed mathematics concepts more than they assessed science concepts. For a number of items, reviewers commented that the item included biology contexts but that the items assessed aspects of experimental design or analysis and that an understanding of biology-specific concepts was not necessary to answer these questions. After coding each assessment form, reviewers were asked to respond to four debriefing questions. Reviewers Comments can be found in Biology **Appendix C** and Debriefing Summary Notes can be found in Biology **Appendix D**.

Reliability among Reviewers
Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in **Table 47**, on the following page, were computed after adjudication. Because of the small number of items, average percent agreement is reported for all items for which a majority of reviewers coded the item to a GSE.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other six reviewers. For most alignment studies, the standards pairwise agreement is higher than 0.60. The pairwise agreement for assigning standards to items was very high for all test forms (0.93 or greater). This is largely because reviewers coded most items to the general science category and for most of these items—although not for all—the reviewers found that the items were very clearly not related to the biology standards (e.g. were contextualized in the physical sciences). For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For all test forms, pairwise agreement for reporting category is very high (0.95 or greater).

Table 47. Intraclass and Pairwise Comparisons, ACT Science with Biology GSE

Test Form	Percent Agreement for DOK (# of items)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
ACT 74C	88% (6)	1.0	0.99
ACT A10	93% (2)	0.95	0.95
ACT 74H	92% (8)	0.95	0.93

Conclusion

The central research question for the alignment analysis was to what degree the ACT or SAT was aligned with the Georgia Standards of Excellence for American Literature and Composition, Algebra I, Geometry, and Biology. The content analysis was conducted to help inform a decision about whether either or both of these nationally recognized assessments could be used in lieu of the Georgia Milestones End-of-Course assessments for American Literature and Composition, Algebra I, Geometry, and Biology.

Study results are based on a content analysis of two ACT and two SAT test forms for ELA and mathematics and three ACT test forms for science. For all content areas and test forms, alignment study results suggest that extensive augmentation to assessments would be required for either the ACT or SAT to meet minimum alignment criteria with the GSE for the corresponding courses. The least augmentation would be needed for alignment with the GSE for American Literature and Composition, with the ACT needing an average of 12 items revised or replaced and the SAT needing around 14 items revised or replaced. For minimum alignment with the Algebra I GSE, the ACT would need 48 items revised or replaced and the SAT would need an average of 34 items revised or replaced. For minimum alignment with the Geometry GSE, the ACT would need around 30 items revised or replaced and the SAT would need 34 items revised or replaced. For Biology, the ACT science test forms would need an average of 33 items revised or replaced. In current form, neither the ACT nor the SAT would yield student scores that could be used to make valid inferences about student proficiency as it relates to the Georgia Standards of Excellence.

References

American Association for the Advancement of Science (1993). *Benchmarks for Science Literacy.* Oxford University Press, New York.

National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III.* Mahwah, NJ: Lawrence Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.