Achieve

Independent Analysis of the Alignment of the ACT to the Common Core State Standards

PREPARED BY: ACHIEVE

MARCH 2018

TABLE OF CONTENTS

Section	Page #
Executive Summary	3
Introduction	8
Study Findings	16
References	34
Appendix A: About Achieve	36
Appendix B: Methodology	37
Appendix C: Changes to Center for Assessment Methodology	58

EXECUTIVE SUMMARY

In 2016, 20 states administered the ACT exam to all high school students. Of those 20 states, three use the ACT in their statewide testing program as an assessment for accountability at the high school level. States using the tests as part of an accountability system are required to submit results from an independent alignment study to the U.S. Department of Education for the federal peer review process. Achieve conducted an independent analysis of the alignment of the ACT plus Writing with the Common Core State Standards (CCSS). This study examines four ACT subtests — English, Reading, Writing, and Mathematics.

Achieve is a nonprofit, nonpartisan education policy organization with a long history of working with states to support the development and implementation of college- and career-ready academic standards and graduation requirements, improve assessments, and strengthen accountability. Achieve has extensive experience conducting standards and assessment reviews for states (see Appendix A for more about Achieve). In conducting this alignment study, Achieve used an evaluation approach adapted from a methodology developed by the National Center for the Improvement of Educational Outcomes. The Center developed the methodology based on a set of criteria, Criteria for Procuring and Evaluating High Quality Assessments (2014), from the Council of Chief State School Officers (CCSSO).

The Achieve review consisted of two major phases. The first phase was an outcomes review of the test items and forms along with a generalizability review using test documentation provided by ACT. In the outcomes review, the reviewers examined the item alignment and quality documentation from the test developer and validated that the items were aligned to the standards as the test developer claimed and met a number of measures of quality. Program documentation was used to determine the generalizability of the findings from the outcomes review to other test forms. The second phase was the development of overall ratings for each criterion. The results of the outcomes reviews of items in phase one were rolled up to the form levels and then to the overall testing program level. Reviewers used their individual ratings to inform consensus discussions to agree on final ratings for each item and criterion.

A summary of Achieve's analysis of the ACT plus Writing is shown below and is described in more detail beginning on page 8. Summary ratings for both English language arts/literacy (ELA) and mathematics have two components:

- Content rating, which combines the scores that are relevant to describing the degree to which the ACT focuses on the content most important for high school as determined by the CCSS in ELA and mathematics;

- Depth rating, which combines appropriate criterion scores to provide an overall description of the degree to which the ACT reflects the depth of the CCSS, including the percentage of items aligned to content standards.

The criteria and methodology are explained in more detail in Appendices B and C.

FINDINGS: ELA

The overall Content rating for the ACT in ELA was rated as a **Weak Match.** While the assessment program emphasizes some aspects of close reading and language skills as well as real-world activities, it fails to require students to use direct textual evidence, demonstrate a range of writing skills, or focus on the vocabulary most appropriate to high school. Moreover, the ACT English, Reading, and Writing subtests do not explicitly measure research or inquiry.

ELA: Overall Content Rating and Criterion Ratings

Criterion	Rating
Overall Content Rating	**Weak Match**
B.3: Requiring students to read closely and use evidence from text	**Weak Match**
B.5: Assessing writing	**Weak Match**
B.6: Emphasizing vocabulary and language skills	**Limited/ Uneven Match**
B.7: Assessing research and inquiry	**Weak Match**

In evaluating the depth of ELA in the ACT, more than 50 percent of items reviewed are not aligned to the claimed ELA content standards. While texts are of high quality, they do not strike an appropriate balance between informational and literary. The range of cognitive demand was not given an overall rating, though reviewers tended to assign Domain of Knowledge (DOK) ratings at a lower level of complexity than ACT's ratings. Overall, the Depth rating was rated as a **Good Match,** though reviewers were concerned about the level of item alignment to the CCSS.

ELA: Overall Depth Rating and Criterion Ratings

Criterion	Rating
Overall Depth Rating	**Good Match**
B.1: Using a balance of high-quality literary and informational texts	**Good Match**
B.2: Focusing on the increasing complexity of texts across grades	**Good Match**
B.4: Requiring a range of cognitive demand	**[Not Rated]**
B.9: Ensuring high-quality items and a variety of item types	**Limited/ Uneven Match**

FINDINGS: MATHEMATICS

The overall Content rating for mathematics on the ACT was rated as a **Weak Match.** While the reviewed forms balance score points across conceptual, procedural, and application items, the assessment does not focus on the content most important for high school as described by the Widely Applicable Prerequisites.[1] Fewer than a third of the items on the ACT Math subtest align to the appropriate content targets for high school, resulting in a Weak Match for Content, as C.1 is heavily weighted in the overall rating.[2]

Mathematics: Overall Content Rating and Criterion Ratings

Criterion	Rating
Overall Content Rating	**Weak Match**
C.1: Focusing strongly on the content most needed for success in later mathematics	**Weak Match**
C.2: Assessing a balance of concepts, procedures, and applications	**Excellent Match**

For Depth, mathematics on the ACT was rated a **Weak Match.** A limited number of items on the assessment (fewer than 50 percent) are aligned to the claimed mathematical content standards. While an appropriate proportion of the assessment reflects mathematical practices, many of those items do not connect to mathematical content standards. The items on the ACT Math subtest are largely well written, with relatively few instances of editorial issues. The range of item types included on the assessment is limited, with all mathematics items consisting of multiple-choice items. Cognitive demand (C.4) was not given a rating, but reviewers agreed with approximately half of the DOK levels claimed by ACT.

Mathematics: Overall Depth Rating and Criterion Ratings

Criterion	Rating
Overall Depth Rating	**Weak Match**
C.3: Connecting practice to content	**Weak Match**
C.4: Requiring a range of cognitive demand	**[Not Rated]**
C.5: Ensuring high-quality items and a variety of item types	**Weak Match**

[1] Ratings were based on the extent to which items on the assessment align exclusively to prerequisites for careers and a wide range of postsecondary studies and to the domains within the Widely Applicable Prerequisites, and they are drawn from the *High School Publishers' Criteria for the Common Core State Standards for Mathematics* (2013).

[2] As described in the *ACT Technical Manual Supplement*, 40–43 percent of ACT mathematics items are intended to measure content below the 8th grade level (Integrating Essential Skills), so the design of ACT inherently limits the number of items intended to measure high school mathematics content.

RECOMMENDATIONS FOR STATES CONSIDERING THE USE OF THE ACT AS A STATEWIDE SUMMATIVE ASSESSMENT

Based on the findings of the study, Achieve developed a set of policy and assessment-specific recommendations for states that would like to use the ACT as their statewide summative assessment. These recommendations are intended to support the improvement of a state's high school assessment program so that it more fully reflects the state's academic standards in ELA and mathematics.

Policy Recommendations

States should not use the ACT as the statewide accountability measure for ELA and mathematics. States are using college admissions tests such as the ACT in school accountability in a couple of appropriate ways — as part of a college readiness indicator or as an equity measure to increase access to college for all students. However, using the ACT as the primary measure of math and ELA achievement of the state college and career readiness standards for accountability is ill advised for both alignment considerations and technical reasons.

States should not allow districts to administer the ACT in lieu of the statewide summative assessment. While the Every Student Succeeds Act gives states the opportunity to provide districts the choice to administer a college admissions test, including the ACT, instead of the statewide summative assessment, states should be cautious about opening this door. Beyond the alignment issues described in this study, ensuring comparability between state assessments and college admissions tests will be extremely challenging, if not impossible, for purposes of accountability. With this choice, districts might "shop around" for the assessment that shows them in the best light, while avoiding tough but necessary conversations about the performance of all students.

States that have adopted the ACT should ask ACT to augment their tests to improve alignment to send educators better signals about instruction. This study recommends that the ACT be augmented with additional items to bring it into better alignment with states that have adopted the CCSS. Augmentation presents challenges, particularly added cost and complexity with such an assessment. However, Achieve urges ACT to respond positively to these recommendations because doing so will send more consistent signals to educators and will increase the likelihood of the ACT meeting peer review guidelines.

Recommendations for ELA

ELA items should fully align to the grade-level standards. Reviewers agreed with fewer than 50 percent of ACT's claimed alignments to the CCSS. To make evidence-based claims that the content standards intended for all students have been met, ACT should ensure that its assessment includes more accurate claims of alignment at the item level.

The assessment should include texts that reflect a balance of literary and informational text types. The CCSS require a 50/50 balance of literary and informational passages by grade, largely based on the 2009 National Assessment of Educational Progress (NAEP) Reading Framework. As such, the passages on the ACT should reflect a more balanced distribution of both literature (stories, drama, and poetry) and informational text, including literary nonfiction.

ACT should publish text complexity information. Text complexity is one metric to ensure that students are prepared to meet postsecondary reading demands and is a priority in the CCSS. Because of its importance,

educators need tools to determine text complexity and examples of texts that are appropriate to the grade. Thus, ACT should publish sample quantitative and qualitative analyses of texts, so educators can have clear examples of texts that not only reflect the standards but also are similar to texts students will encounter on the assessment.

The assessment should require close reading and the use of evidence. The CCSS directly require that students "read closely to determine what the text says … [and] cite specific textual evidence when writing or speaking to support conclusions drawn from the text." More items on the ACT should require close reading and the use of evidence from text to meet the requirements of the standard.

The assessment should require students to draw upon text(s) to craft their writing and vary the types writing that are assessed. Writing standards 1 and 2 require students to use relevant and sufficient evidence and convey information through the effective selection, organization, and analysis of content. Therefore, students need to use evidence and analyze content in their writing; that is, students need to encounter rich texts on the assessment that can support writing to sources. Additionally, the CCSS require a distribution of writing purposes (to persuade, to explain, to convey experience) as reflected in the 2011 NAEP Writing Framework. To meet this requirement, the ACT should require students to write in a variety of genres (argument, informative/explanatory, narrative) to meet the standards.

Recommendations for Mathematics

Math items should fully align to the grade-level standards. Reviewers found fewer than 50 percent of claimed alignments to the CCSS. To make evidence-based claims that the content standards intended for all students have been met, ACT should ensure that its assessment includes more accurate claims of alignment at the item level.

More than 50 percent of mathematics items should align to the Widely Applicable Prerequisites. Additionally, assessment documentation should indicate alignment to the progressions documents (Common Core Standards Writing Team, 2013; National Governors Association et al, 2013).

The assessment should ensure that the Standards for Mathematical Practice are included in assessment design. These practices are a component of student proficiency and are included in the CCSS.

INTRODUCTION

In 2016, 20 states administered the ACT exam to all high school students. Of those 20 states, three use the ACT in their statewide testing program as an assessment for accountability at the high school level. States using the tests as part of an accountability system are required to submit results from an independent alignment study to the U.S. Department of Education for the federal peer review process. Achieve conducted an independent analysis of the alignment of the ACT plus Writing with the Common Core State Standards (CCSS).

Achieve is a nonprofit, nonpartisan education policy organization with a long history of working with states to support the development and implementation of college- and career-ready academic standards and graduation requirements, improve assessments, and strengthen accountability. Achieve has extensive experience conducting standards and assessment reviews for states (see Appendix A for more about Achieve). In conducting this alignment study, Achieve used an evaluation approach adapted from a methodology developed by the National Center for the Improvement of Educational Outcomes. The Center developed the methodology based on a set of criteria, *Criteria for Procuring and Evaluating High Quality Assessments* (2014), which was developed by the Council of Chief State School Officers (CCSSO). The criteria reflect best practices for assessment development and are grounded in the research that defines college and career readiness in mathematics and English language arts/literacy (ELA). This methodology was operationalized and used in a set of studies by the Thomas B. Fordham Institute and HumRRO to examine the depth and breadth of several widely used and well-known assessment programs — ACT Aspire, the Massachusetts Comprehensive Assessment System (MCAS), the Partnership for Assessment of Readiness for College and Careers (PARCC), and the Smarter Balanced Assessment Consortium. For this study Achieve used a similar process, with modifications based on recommendations from the Fordham and HumRRO reports as well as other changes identified by Achieve, external experts, and the study's Technical Advisory Committee.

The ACT

The ACT is a national college admissions exam that consists of four sections: English, Reading, Mathematics, and Science, with an optional Writing test. The ACT has long been recognized as a predictor of college and career readiness. The test is developed using ACT's own ACT College and Career Readiness Standards, which the organization says are also aligned to many states' college- and career-ready standards (ACT, 2015). Additionally, ACT, along with Achieve and the College Board, formed the writing team for the development of the college and career readiness Anchor Standards, which preceded the development of the CCSS, and ACT endorsed those standards (ACT, 2010b). Today, ACT claims that its tests align to state college- and career-ready standards, including the CCSS.

This study examines four ACT subtests — English, Reading, Writing, and Mathematics. Below, each of the ACT subtests is briefly described. The information below has been adapted from the *ACT Technical Manual* (ACT, 2010a) and *ACT Technical Manual Supplement* (ACT, 2016).

English, Reading, and Writing

- The ACT English test consists of 75 multiple-choice items, and there are four reported scores: an overall score, Production of Writing, Knowledge of Language, and Conventions of Standard English.

- Production of Writing: Students apply their understanding of the rhetorical purpose and focus of a piece of writing to develop a topic effectively and use various strategies to achieve logical organization, topical unity, and general cohesion.

- Knowledge of Language: Students demonstrate effective language use through ensuring precision and concision in word choice and maintaining consistency in style and tone.

- Conventions of Standard English: Students apply an understanding of the conventions of standard English grammar, usage, and mechanics to revise and edit text.

The ACT Reading test consists of 40 questions, with five reported scores — a total test score, three reporting category scores based on specific language and skills (Key Ideas and Details, Craft and Structure, and Integration of Knowledge and Ideas), and an Understanding Complex Texts indicator. The three reporting category scores based on specific language and skills are described briefly below.

- Key Ideas and Details: Students read texts closely to determine central ideas and themes; summarize information and ideas accurately; and read closely to understand relationships and draw logical inferences and conclusions, including understanding sequential, comparative, and cause-effect relationships.

- Craft and Structure: Students determine word and phrase meanings, analyze an author's word choice rhetorically, analyze text structure, understand authorial purpose and perspective, and analyze characters' points of view. They interpret authorial decisions rhetorically and differentiate among various perspectives and sources of information.

- Integration of Knowledge and Ideas: Students understand authors' claims, differentiate between facts and opinions, and use evidence to make connections among different texts that are related by topic. Some questions require students to analyze how authors construct arguments, evaluating reasoning and evidence from various sources.

The ACT Writing test is an optional 40-minute essay that was revised in 2015. The one writing prompt describes a complex issue and presents three different perspectives on that issue. Students are asked to read the prompt and write an essay in which they develop their own perspective on the issue. The essay must analyze the relationship between their own perspective and one or more other perspectives. Students receive five scores in Writing: a total writing score (ranging from 2 to 12 points) and four domain scores, also 2–12, based on an analytic scoring rubric. The subject-level score is the rounded average of the four domain scores:

- Ideas and Analysis;

- Development and Support;

- Organization; and

- Language Use and Conventions.

Mathematics

The Mathematics test contains 60 multiple-choice items in three major reporting categories: Preparing for Higher Math, Integrating Essential Skills, and Modeling. Preparing for Higher Math includes five reported subscores — Number and Quantity, Algebra, Functions, Geometry, and Statistics and Probability.

- Preparing for Higher Math: This reporting category captures the more recent mathematics that students are learning, starting when students begin using algebra as a general way of expressing and solving equations.

- Integrating Essential Skills: This reporting category includes mathematics learned before 8th grade, including rates and percentages; proportional relationships; area, surface area, and volume; average and median; expressing numbers in different ways; using expressions to represent quantities and equations to capture relationships; and other topics.

- Modeling: This reporting category includes all questions that involve producing, interpreting, understanding, evaluating, and improving models. Each modeling question is also counted in the other two reporting categories above.

Methodology

In 2013, leading educational researchers and measurement experts responded to the widespread adoption of college- and career-ready standards among states by releasing a set of research-based criteria for high-quality assessments that would support measuring student learning against these new standards (Darling-Hammond et al., 2013). These criteria included assessing higher-order thinking skills, assessing critical abilities with high fidelity, and assessing against internationally benchmarked standards. CCSSO transformed these insights into *High-Quality Summative Assessment Principles for ELA/Literacy and Mathematics Assessments Aligned to College and Career Readiness Standards* (2013) and then into a more formalized set of criteria, *Criteria for Procuring and Evaluating High Quality Assessments* (2014). CCSSO's criteria are grouped into five main categories:

A. Meet Overall Assessment Goals and Ensure Technical Quality: This criterion focuses on ensuring that the scores and performance levels indicate progress toward college and career readiness, the assessment is valid for intended purposes, the assessment is reliable, the assessment is designed to yield valid and consistent test interpretations within and across years, accessibility is provided to all students, test design and interpretations are transparent, and the assessment meets all requirements for data privacy and ownership.
B. *Align to Standards — ELA (see below)*
C. *Align to Standards — Mathematics (see below)*
D. Yield Valuable Reports on Student Progress and Performance: This criterion calls for evidence that score reports focus on student achievement and progress to readiness and that they provide timely data that inform instruction.
E. Adhere to Best Practices in Test Administration: This criterion calls for evidence that assessment systems maintain necessary standardization and ensure test security.
F. State-Specific Criteria: Evidence for this criterion will vary by state but may include the involvement of multiple stakeholder groups in designing, developing, and scoring assessments; procuring a system of aligned assessments; or ensuring the interoperability of computer-based items.

The National Center for the Improvement of Educational Assessment (also known as the Center for Assessment) developed a methodology based on a subset of CCSSO's 2014 criteria, focusing on the criteria for ELA and mathematics (categories B and C in the above list) (2016). The CCSSO criteria for ELA and mathematics specifically focus on the highest priority knowledge and skills at each grade band described by college- and career-ready standards, such as the CCSS. By using the CCSSO criteria as the focus of the alignment methodology, test evaluations can provide useful information about the degree to which the tests focus on the essential knowledge and skills described by the standards and give clear signals about instructional priorities at the targeted grade level.

Criteria are organized into two categories in each content area: Content and Depth. The Content ratings reflect the degree to which the assessment focuses on the highest priority content within each subject area; the Depth ratings reflect the degree to which assessments measure the depth and complexity reflected by college- and career-ready standards. The Center for Assessment also provided scoring guidance and tentative cutoffs within each criterion, yielding ratings described as Excellent, Good, Limited/Uneven, or Weak Match.

The *CCSSO Criteria for Procuring and Evaluating High Quality Assessments* served as the foundation for Achieve's methodology, as well as the methodology used by the Fordham Institute and HumRRO in their grades 3–8 and high school reviews of several statewide summative assessments — ACT Aspire, MCAS, PARCC, and Smarter Balanced (Doorey & Polikoff, 2016; Schultz, Michaels, Dvorak, & Wiley, 2016, respectively). Achieve's approach, adapted from the Center for Assessment's methodology, similarly focused on the CCSSO alignment criteria for ELA and mathematics, shown below in Table 1.

Table 1. CCSSO Alignment Criteria for ELA and Mathematics

Align to Standards: ELA	
Test Content Criteria	**Test Depth Criteria**
● B.3: Requiring students to read closely and use evidence from text	● B.1: Assessing student reading and writing achievement in both ELA and literacy[4]
● B.5: Assessing writing	● B.2: Focusing on complexity of texts[5]
● B.6: Emphasizing vocabulary and language skills	● B.4: Requiring a range of cognitive demand
● B.7: Assessing research and inquiry	● B.9: Ensuring high-quality items and a variety of item types
● B.8: Assessing speaking and listening[3]	

[3] This review does not evaluate B.8: Assessing speaking and listening because in feedback from the U.S. Department of Education on its initial peer review submission, the state requesting the review received a speaking and listening waiver and was not expected to submit additional evidence regarding speaking and listening during the period of the waiver.
[4] The title of this criterion was changed to "Using a balance of high-quality literary and informational texts" for this study to better reflect its content.
[5] The title of this criterion was changed to "Focusing on the increasing complexity of texts across grades" to better reflect its content.

Align to Standards: Mathematics	
Test Content Criteria	**Test Depth Criteria**
● C.1: Focusing strongly on the content most needed for success in later mathematics ● C.2: Assessing a balance of concepts, procedures, and applications	● C.3: Connecting practice to content ● C.4: Requiring a range of cognitive demand ● C.5: Ensuring high-quality items and a variety of item types

A brief description of each criterion is provided below.

ELA

B.1: Using a balance of high-quality literary and informational texts: The focus of this criterion is on the balance and quality of the texts students read as part of the assessment. Subcriteria within B.1 relate to the major focus and advances associated with college- and career-ready ELA standards, including the CCSS: the balance of informational and literary texts, the quality of the passages included, and the key features and types of informational texts students read and to which they respond.

B.2: Focusing on the increasing complexity of texts across grades: The focus of this criterion is on text complexity: whether the passages on the test forms reviewed are of appropriate complexity for the grade and whether the documentation indicates extensive qualitative and quantitative measures used to determine the appropriate complexity for a given grade.

B.3: Requiring students to read closely and use evidence from text: The focus of this criterion is to highlight the important elements of reading as described by college- and career-ready standards like the CCSS. The subcriteria associated with B.3 emphasize the use of direct textual evidence, close reading, and focusing on central ideas.

B.4: Requiring a range of cognitive demand: The focus of this criterion is to determine whether the assessment items on a given test form, as well as the assessment program as a whole, reflect a range of cognitive demand that is sufficient to assess the depth and complexity of the state's standards, as evidenced by use of a generic taxonomy (e.g., Webb's Depth of Knowledge [DOK]) or, preferably, classifications specific to the discipline and drawn from mathematical factors.

B.5: Assessing writing: This criterion focuses on writing prompts that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities. The subcriteria that comprise B.5 focus on a balance of writing types across exposition, argument, and narrative types and the extent to which writing prompts require students to write to sources.

B.6: Emphasizing vocabulary and language skills: This criterion focuses on assessing proficiency in the use of language, including vocabulary and conventions that reflect college and career readiness. The eight subcriteria associated with B.6 accordingly focus on features of the vocabulary and language items and the percentage of score points associated with each on an assessment.

B.7: Assessing research and inquiry: This criterion focuses on whether assessments ask students to

demonstrate research and inquiry skills, as demonstrated by the ability to find, process, synthesize, organize, and use information from sources.

B.9: Ensuring high-quality items and a variety of item types: This criterion focuses on three major aspects of items within an assessment: the technical and editorial quality of items, the variety of item types on forms, and the alignment of items to standards.

Mathematics

C.1: Focusing strongly on the content most needed for success in later mathematics: This criterion focuses on college and career readiness by requiring that the vast majority of the items and score points on an assessment focus on the content most needed for later success in mathematics. In high school specifically, meeting this criterion means that at least half of the points in each grade/course align exclusively to the prerequisites for careers and a wide range of postsecondary studies, as described in the Widely Applicable Prerequisites document (National Governors Association et al, 2013).

C.2: Assessing a balance of concepts, procedures, and applications: This criterion focuses on determining whether an assessment measures conceptual understanding, fluency and procedural skills, and application of mathematics as set out in college- and career-ready standards.

C.3: Connecting practice to content: This criterion focuses on the integration of Standards for Mathematical Practice with content standards by requiring that assessments connect the most important mathematical content of the grade or course to mathematical practices (e.g., modeling and making mathematical arguments).

C.4: Requiring a range of cognitive demand: This criterion focuses on ensuring that assessments require all students to demonstrate a range of higher-order, analytical thinking skills in math based on the depth and complexity of college- and career-ready standards.

C.5: Ensuring high-quality items and a variety of item types: This criterion focuses on ensuring that high-quality items and a variety of item types are strategically used to appropriately assess the standard(s). This criterion focuses on three major aspects of items within an assessment: the technical and editorial quality of items, the variety of item types on forms, and the alignment of items to standards.

ACHIEVE'S REVIEW OF THE ACT

Prior to engaging in this review, Achieve revised the methodology developed by the Center for Assessment and used by the Fordham Institute and HumRRO based on recommendations in those studies as well as feedback from review authors and experts in the field. The Achieve methodology and scoring guidance can be found in Appendix B, and significant changes to the original methodology developed by the Center for Assessment can be found in Appendix C.

The Achieve review consisted of two major phases, both of which are described in more detail below. The first phase was an outcomes review of the test items and forms along with a generalizability review using test documentation provided by ACT. In the outcomes review, the reviewers examined the item alignment and documentation from the test developer and validated that the items were aligned to the standards as claimed and met a number of measures of quality. Program documentation was used to determine the generalizability of the findings from the outcomes review to other test forms. The second phase was the development of overall ratings for each criterion through reviewer consensus meetings. Reviewers used their individual ratings to inform consensus conversations with other reviewers. Final determinations were made, by consensus, for each criterion on each individual item.

Phase 1: Outcomes and Generalizability Review

The outcomes review required individual, independent evaluation of actual student test items by expert content reviewers against each of the subject area criteria. ACT provided two complete, recently administered test forms (pseudonymously referred to as Form 1 and Form 2 throughout this report), test item metadata, and descriptive data about individual test items and passages. This information included alignment to the CCSS as determined by ACT, the level of text complexity for reading passages, cognitive complexity information (in this case, DOK ratings for each item), and item type (e.g., multiple choice or constructed response). Reviewers rated items for each criterion based on the documentation provided and on their expert judgment.

Reviewers also independently evaluated documentation provided by ACT to determine whether the results from reviewing the two ACT test forms were generalizable to the test program as a whole. ACT provided both public-facing and confidential documentation to support the generalizability review, including technical manuals and reports, score reporting guidelines, item development guidelines, scoring rubrics (e.g., for writing), style guides, and task models. These materials detailed ACT's approach to building and validating its assessments to demonstrate that they adequately assess college and career readiness and the CCSS.

Phase 2: Development of Final Ratings

Following the independent reviewer ratings, conducted over a two-week period, reviewers met in person to discuss their individual scores and comments in consensus meetings. The reviewers came to consensus on final scores for items on each form, reflecting the collective judgment of the panel. Next, the quantitative scores associated with individual items were aggregated into subcriterion-level scores, which provided a tentative rating for each subcriterion. Review panels used the tentative ratings and their expert judgment to determine final outcomes for subcriterion ratings, taking into account the scores across both ACT forms that were evaluated. Reviewers then considered the results of the generalizability review and came to consensus at the item level on the degree to which the assessment reflected each respective criterion. They issued final ratings

at the criterion level on the following scale: Excellent, Good, Limited/Uneven, or Weak Match to the criterion. These final criterion ratings were weighted to produce aggregate Content and Depth scores for each content area. Reviewers also developed summary statements with a rationale for the ratings as well as the observed strengths and areas of improvement for the ACT.

Reviewers and Training

In keeping with Achieve's historical approach to assessment and standards reviews, this review was conducted using small teams of Achieve and external content-level experts and was supported by Achieve staff. Each content area — ELA and mathematics — used three expert reviewers (one Achieve content lead and two external reviewers). The Center for Assessment (2016) noted that "the quality of an assessment evaluation depends in large part on the selection of qualified and experienced yet impartial reviewers" (p. 11) and highlighted deep content knowledge; classroom content teaching experience; knowledge and familiarity with college and career readiness standards; knowledge of large-scale assessment specifications and information; and familiarity with test items, performance tasks, task design templates, and scoring guides. For this review, external reviewers were selected based on those criteria.

Reviewer training took place in three phases in June and July 2017. First, reviewers participated in training modules that outlined the study, each of the CCSSO criteria, and approaches to scoring for both the generalizability and outcomes ratings. Second, once initially trained, reviewers calibrated their scoring approach using a complete test form of a large-scale high school assessment designed to measure college and career readiness. Third, reviewers participated in an interactive presentation from ACT that described the components of the assessment and the materials that reviewers would see during the review. Immediately following the ACT presentation and calibration exercises, reviewers had access to the full suite of ACT materials to begin the study.

Reviewer Consensus Model

The approach to generating item and final ratings used in the Achieve review differed from both traditional alignment studies and the Center for Assessment's methodology as used by Fordham and HumRRO. Approaches such as Webb's generally use larger panels of teachers and other experts to independently judge how items are aligned to standards and measure cognitive complexity (generally using DOK). These data are used to generate indices of agreement to the test developer's alignment and cognitive complexity claims, as well as inter-rater reliability coefficients. The Fordham/HumRRO approach used small panels of reviewers to independently rate test forms and generalizability evidence across the range of CCSSO criteria, which generated tentative scores that reviewers used to generate consensus ratings at the subcriterion level. For Achieve's review, reviewers independently reviewed and scored test forms and generalizability evidence, which generated tentative scores that were brought into the consensus meetings. At the consensus meetings, reviewers used their own scores and evidence in discussion to come to consensus on individual items and generalizability evidence across all criteria. These consensus scores were then used to generate final ratings for outcomes and generalizability. This approach is in keeping with the design of the methodology by the Center for Assessment. As the Center for Assessment (2016) described, "Reviewers engage in a process of discussion and consensus building, to move from reviewer-level results of a single test form to panel-level results across a testing program" (p. 16).

STUDY FINDINGS

ELA: Content Criteria

Achieve's analysis finds that the ACT English, Reading, and Writing subtests do not strongly emphasize content in ELA as defined by the expectations in the CCSS. The assessment program emphasizes some aspects of close reading and language skills, as well as real-world activities, but fails to require students to use direct textual evidence, demonstrate a range of writing skills, or focus on the vocabulary most appropriate to high school. Moreover, the ACT English, Reading, and Writing subtests do not measure research or inquiry. Summary ratings are presented in the chart below, followed by a discussion of the components that make up the Content rating (B.3, B.5, B.6, and B.7).

ELA: Overall Content Rating and Criterion Ratings

Criterion	Rating
B.3: Requiring students to read closely and use evidence from text Although most items focus on the central idea and important particulars of the text and require close reading of some kind, too many items can be answered by simply matching language from answer choices to the text without requiring students to read closely. Moreover, no items require direct textual evidence in support of a conclusion, generalization, or inference drawn from the text.	**Weak Match**
B.5: Assessing writing Only a single type of writing (argumentative) is assessed on both forms, providing insufficient opportunity to assess writing of multiple types. Moreover, the writing prompts do not require students to use direct textual evidence from sources or draw inferences from texts.	**Weak Match**
B.6: Emphasizing vocabulary and language skills The test contains a sufficient number of high-quality language use items, but items coded as testing vocabulary are more uneven. The test forms contain relatively few vocabulary items/score points, and many items coded to vocabulary do not test Tier 2 words or words or phrases central to the text. They also do not require students to use context to determine the meaning of words. Test documentation supports this finding, indicating a great deal of emphasis on vocabulary across the Reading and English subtests but limited focus on Tier 2 words or using context to construct the meaning of words.	**Limited/ Uneven Match**
B.7: Assessing research and inquiry The assessment has no items devoted to research.	**Weak Match**

Criterion B.3: Requiring Students to Read Closely and Use Evidence from Text

The CCSS emphasize that students should read narrative and informational text carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct answers. Test questions that require close reading will ask students to uncover layers of meaning that lead to a deep comprehension of the overall text and a full understanding of central ideas and important particulars. These items ask for

a careful analysis of textual elements (e.g., central ideas and their development, claims and supporting evidence, characterization, plot development) that contribute to the meaning of the text. To fully meet this criterion in this study, nearly all items on reviewed test forms must require close reading and analysis of text, focus on central ideas and important particulars, and be aligned to the specifics of the grade-level standards. Additionally, more than half the reading score points should be based on items requiring direct use of textual evidence. These criteria should be supported by additional generalizability evidence.

Overall, reviewers determined a Weak Match to the criteria requiring students to read closely and use evidence from text. In their review, they found that just 12 percent of Reading items align to the assigned CCSS grade-level standards. The breakdown by expectation and form is outlined in Table 2 below. Reviewers found that 56 percent of Reading items across both forms require close reading and analysis of text. While many Reading items ask students to return to the reading passage, the purpose is to find and match similar language between the stem and the correct answers rather than close reading and analysis of text. Reviewers determined that 79 percent of Reading items across both test forms ask students to focus on central ideas or important particulars of the text. No items were found to ask students to use direct textual evidence to support a claim or inference about the text.

Table 2. Percentage of Items Meeting Subcriteria for B.3: Requiring Students to Read Closely and Use Evidence from Text

B.3 Subcriterion	Overall	Form 1	Form 2
Alignment to grade-level reading standards	12%	13%	10%
Close reading and analysis of text	56%	53%	58%
Focusing on central ideas and important particulars	79%	88%	70%
Direct use of textual evidence	0%	0%	0%

Criterion B.5: Assessing Writing

The CCSS emphasize writing tasks that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities. To fully meet this criterion, expository and argumentative writing types should be balanced across all forms, and writing prompts should require students to confront text or other stimuli directly, draw on textual evidence, and support valid inferences from text or stimuli. Both the distribution of writing types and the requirements for writing prompts should be thoroughly documented in the provided generalizability information. Importantly, the CCSS emphasize that students must demonstrate their facility with the writing standards directly through the production of writing.

The ACT codes two types of items to writing standards across the English, Reading, and Writing subtests:

- Standalone multiple-choice language items, which require students to read another commissioned text and respond to multiple-choice tasks but not to produce writing; and

- The argumentative writing prompt, which requires the production of writing.

Due to the absence of balance in the types of writing students are asked to demonstrate (i.e., a single writing prompt of the same type on both Forms 1 and 2), the limited use of text to support the writing students generate, and the use of multiple-choice items that require no writing, the ACT is Weak in its assessment of the writing expectations described by the CCSS.

No items on the ACT English tests coded by ACT to writing standards were found to meet the expectations for writing as articulated in the CCSS, as none of these items require writing of expository or argumentative pieces. Additionally, the documentation provided indicates that only argumentative writing is assessed. To fully meet this criterion, students must also be asked to draw on textual evidence to support claims or to support valid inferences from the text. While the argumentative writing prompts in both Writing subtest forms require students to confront text directly by considering a brief introduction and three short commissioned texts (ranging from 15 to 95 words each), there was not enough text for students to draw sufficient evidence to develop or support a claim. The writing prompts require students to confront text directly, but they do not require "writing to sources," as students must use their own ideas as well as draw on the very limited amount of text provided. The text is not sufficient for students to support valid inferences from text or stimuli.

Criterion B.6: Emphasizing Vocabulary and Language Skills

Assessments measuring the CCSS should require students to demonstrate proficiency in the use of language, including vocabulary and conventions. This criterion evaluates both vocabulary and language items. To fully meet this criterion, the large majority (75 percent or more) of vocabulary items should reflect the requirements for college and career readiness, including focusing on general academic (Tier 2) words, asking students to use context to determine word meaning, and assessing words that are important to the central ideas of the text. Similarly, a large majority (75 percent or more) of the items in the language skills component and/or scored with a writing rubric should mirror real-life activities, focus on common errors, and emphasize the conventions most important for readiness. Documentation should thoroughly outline the requirement that language is assessed in the writing rubric or that language skills mirror real-world activities, focus on common student errors, and emphasize the conventions most important for readiness. Finally, both vocabulary and language should either be reported at a statistically reliable subscore or each comprise at least 13 percent of score points.

The ACT is an Uneven Match to this criterion. The test contains a sufficient number of high-quality language use items, but items coded as testing vocabulary are more uneven. The test forms contain relatively few vocabulary items/score points, and many items coded to vocabulary do not test Tier 2 words or words or phrases central to the text. The test also does not require students to use context to determine the meaning of words. Test documentation supports this finding, indicating a great deal of emphasis on vocabulary across the Reading and English subtests but limited focus on Tier 2 words or using context to construct the meaning of words.

Vocabulary

Vocabulary items are found on both the Reading and English portions of the ACT. Across English and Reading, 38 percent of vocabulary items test Tier 2 vocabulary words and require the use of context to determine meaning (29 percent on Form 1 and 47 percent on Form 2). Analysis of the vocabulary items specifically on the Reading subtest revealed that the majority of vocabulary items assess Tier 2 words and words important to understanding the central ideas of the text; fewer vocabulary items on the English test assess Tier 2 words and words important to understanding the central ideas of the text. While documentation provided by ACT indicates that vocabulary is extensively assessed, it does not indicate a focus on Tier 2 vocabulary words or words important to understanding the central ideas of the text.

Language

On the ACT, language skills are evaluated with both the writing task (as indicated by the rubric) and a series of multiple-choice items. Considering the constraints of a timed assessment, many of the multiple-choice language items mirror real-world activities and highlight common student errors. In the reviewed test forms students are asked to edit grammatical and syntactical errors in text. The majority of items reflect skills necessary for readiness, as indicated by the Language Progressive Skills Chart for the CCSS.[6]

Table 3. Percentage of Items Emphasizing Vocabulary and Language Skills (B.6)

B.6	Overall % of items coded to language standards	% of language items on Form 1	% of language items on Form 2
Mirror real-world activities	64%	85%	42%
Test conventions most important for readiness	55%	73%	36%
Test common student errors	52%	71%	33%
Mirror real-world activities, test conventions most important for readiness, and test common student errors	46%	61%	31%

Criterion B.7: Assessing Research and Inquiry

High-quality assessments require students to demonstrate research and inquiry skills through their ability to find, process, synthesize, organize, and use information from sources. Research is addressed in Writing standards 7, 8, and 9 of the CCSS. These standards are categorized in the Writing Anchor Standards under "Research to Build and Present Knowledge" (p. 67). To fully meet this criterion, 75 percent or more of the

[6] www.corestandards.org/ELA-Literacy/L/language-progressive-skills

research tasks should require analysis, synthesis, and/or organization of information; the program must either report a research score or demonstrate that research is significant in another way (e.g., number of score points); and documentation should thoroughly outline the requirement that research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source and often from sources in diverse formats.

The documentation provided does not indicate any emphasis on research. ACT did not code any items on the provided test forms to the research standards. Moreover, reviewers did not identify any items across either form that they would recommend aligning to a research standard. Thus, reviewers assigned a Weak Match for criterion B.7.

ELA: Depth Criteria

Achieve's analysis finds that the ACT English, Reading, and Writing subtests are overall a Good Match to the criteria on Depth, which describes students' access to complex texts, the level of cognitive complexity of individual items, and the quality of items that reflect a variety of item types. The items on the ACT English, Reading, and Writing subtests are based on high-quality texts that are appropriate for high school, and the items are generally free from editorial errors. However, more than 50 percent of items were found to not align to the CCSS, weakening the overall item quality across test forms.

ELA: Overall Depth Rating and Criterion Ratings

Criterion	Rating
B.1: Using a balance of high-quality literary and informational texts The texts are of high quality. However, they do not strike an appropriate balance between informational and literary because all texts included on the forms reviewed for this study are informational.	**Good Match**
B.2: Focusing on the Increasing Complexity of Texts across Grades Test documentation indicates that texts are appropriately complex for the high school level and are placed in the correct grade band based on qualitative and quantitative data. While quantitative complexity scores were provided, information about the qualitative analyses that were performed was insufficient.	**Good Match**
B.4: Requiring a range of cognitive demand Ratings for items on the English, Reading, and Writing tests range from DOK 1 to 3. Reviewers agreed with the assigned DOK for 60 percent of the items across all three subtests; when reviewers disagreed with the assigned DOK, they almost always assigned a DOK within one step of the indicated DOK (e.g., 2 vs. 1 or 2 vs. 3).	**[Not Rated]**
B.9: Ensuring high-quality items and a variety of item types The ACT items are of exceptional technical quality, with few editorial or accuracy errors. While skewed heavily toward multiple-choice items, multiple item formats are used, including a student-generated response item. However, a limited number of items (fewer than 50 percent) are aligned to the CCSS, weakening the overall quality of the items.	**Limited/ Uneven Match**

Criterion B.1: Using a balance of high-quality literary and informational texts

Reading passages that are of high quality and that balance literary and informational genres are a priority in the CCSS. To fully meet this criterion, approximately two-thirds of the texts need to be informational; all or nearly all passages need to be of high quality (i.e., they must be content rich, exhibit exceptional craft and thought, and/or provide useful information); and nearly all informational passages must be expository in structure with the informational texts split nearly evenly among literary nonfiction, history/social science, and science/technical. All of these subcriteria should be supported by documentation from the developer.

The ACT was found to be a Good Match for assessing student reading and writing achievement in ELA and literacy. Reviewers evaluated only the texts that were coded to reading standards (i.e., the texts on the Reading test of the ACT battery, not the texts on the English or Writing subtests). All the passages on the Reading test were previously published and are of high quality. Generally, the informational passages have expository structure. Informational texts are well divided among the literary nonfiction, history/social science, and science/technical genres.

Reviewers noted that the reviewed literary nonfiction passages were misidentified as literature rather than informational texts. Although the CCSS define literary nonfiction as informational texts, literary nonfiction passages are coded as "literary narrative," and the items are aligned to Reading for Literature standards instead of standards for Reading Informational Text. The result of coding the literary nonfiction passages as literary texts is that all 10 passages on the Reading subtests in ACT Forms 1 and 2 are informational texts; no literary texts are used. Documentation indicates that the use of literature is permissible in the "Literary Narrative" category, but none is present in the forms evaluated.

Criterion B.2: Focusing on the Increasing Complexity of Texts across Grades

The CCSS establish the expectation that students will read texts that grow in complexity through the end of 12th grade. To bridge the gap between the complexity of texts students read in high school and those they will encounter in college or the workplace, students need to engage with texts that are structurally, thematically, and linguistically complex for the appropriate grade level. To fully meet this criterion, all or nearly all passages should be placed at a grade band and grade level justified by complexity data. The documentation should thoroughly explain how quantitative data and qualitative analyses are used to ensure that texts are placed at the appropriate grade band and grade level, and the documentation should outline that text complexity increases by grade level across all years of the assessment program, meeting college- and career-ready levels by the end of high school.

Text selection is one of the strongest aspects of the ACT, and the ACT is a Good Match to the criteria on focusing on the increasing complexity of texts across grades. From the documentation provided, reviewers determined that 80 percent of passages on both forms are placed in the correct grade band based on quantitative and qualitative data and that ACT has a quality control process to ensure appropriate text selection. The passages evaluated confirmed that the majority of texts are of the appropriate quantitative complexity, but evidence about the qualitative analyses associated with the texts included on both test forms was insufficient. Although reviewers did not have qualitative analyses for the texts on the forms reviewed, they used their professional judgment and determined that the texts are appropriately complex for grade 11.

Criterion B.4: Requiring a Range of Cognitive Demand

Assessments should require all students to demonstrate a range of higher-order, analytical thinking skills based on the depth and complexity of college- and career-ready standards. A common approach to measuring the cognitive demand of individual assessment items or tasks is the use of Webb's DOK. As developed by Norman Webb (2007), there are four DOK levels, which are briefly described below for ELA.

DOK 1: Students are required to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text, as well as basic comprehension of a text, is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from the text, slight paraphrasing of specific details from the text, or simple understanding of a single word or phrase.

DOK 2: Engagement of some mental processing beyond recalling or reproducing a response is required. Both comprehension and subsequent processing of the text or portions of the text are required. Inter-sentence analysis or inference is required. Some important concepts are covered but not in a complex way.

DOK 3: 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. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge.

DOK 4: Higher-order thinking is central, and knowledge is deep at Level 4. The standard or assessment item at this level will probably be an extended activity, with extended time provided for completing it. 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. Students take information from at least one passage of a text 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.

Achieve did not determine a rating for C.4 because the methodology did not assign a DOK rating for content standards. However, Achieve did calculate the percentage of reviewer agreement with ACT DOK claims for each item and compared these findings with ACT's claimed distribution of DOK as found in the *ACT Technical Manual Supplement*. For this criterion, reviewers independently reviewed each item and then came to consensus on a rating for each item, without knowledge of the developer-claimed rating. On Form 1, reviewers rated an item DOK as higher than the claimed DOK for two items (one English item and one Reading item) and rated 38 items lower than the claimed DOK. On Form 2, reviewers rated an item higher than the claim for four items and placed 48 items one level lower. In three instances across both test forms, reviewers rated an item DOK as two levels below the claimed DOK (i.e., reviewers rated an item as DOK 1 while the claimed level was DOK 3).

Tables 4, 5, and 6 below show reviewer ratings by DOK level for the three subtests included in the ELA rating – Reading, English, and Writing – as well as the target DOK percentages included in the ACT Technical Manual Supplement.

Table 4. ACT Claims and Achieve Reviewer Ratings by DOK Level for the Reading Subtest[7]

	DOK 1	DOK 2	DOK 3
Target as claimed by ACT	13–25%	13–25%	38–63%
Overall (claimed)	27%	37%	34%
Overall (reviewed)	34%	61%	5%
Form 1 (claimed)	33%	33%	33%
Form 1 (reviewed)	35%	58%	5%
Form 2 (claimed)	20%	45%	35%
Form 2 (reviewed)	33%	63%	5%

Table 5. ACT Claims and Achieve Reviewer Ratings by DOK Level for the English Subtest[8]

	DOK 1	DOK 2	DOK 3
Target as claimed by ACT	12–15%	53–60%	26–34%
Overall (claimed)	30%	28%	43%
Overall (reviewed)	30%	61%	9%
Form 1 (claimed)	31%	28%	41%
Form 1 (reviewed)	33%	53%	13%
Form 2 (claimed)	28%	27%	45%
Form 2 (reviewed)	27%	68%	5%

Table 6. ACT Claims and Achieve Reviewer Ratings by DOK Level for the Writing Subtest[9]

	DOK 1	DOK 2	DOK 3
Target as claimed by ACT	0%	0%	100%
Overall (claimed)	0%	0%	100%
Overall (reviewed)	0%	0%	100%
Form 1 (claimed)	0%	0%	100%
Form 1 (reviewed)	0%	0%	100%
Form 2 (claimed)	0%	0%	100%
Form 2 (reviewed)	0%	0%	100%

[7] Percentages may not sum to 100 due to rounding.
[8] Percentages may not sum to 100 due to rounding.
[9] Percentages may not sum to 100 due to rounding.

Criterion B.9: Ensuring High-Quality Items and a Variety of Item Types

High-quality and aligned items — items that are free of technical or editorial flaws and meet the demands of the standards — are essential. It is important that students can take an assessment without confusion over erroneous syntactical or grammatical constructions or confronting inaccurate information. High-quality items should also be accurately aligned to standards. This criterion highlights the importance of a variety of item types to assess the complexity and depth of the standards. To fully meet this criterion, at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., constructed-response extended writing), and all or nearly all operational items reviewed should be accurately aligned to standards and not have technical or editorial flaws.

For this criterion, the ACT English, Reading, and Writing subtests were rated highly on the subcriteria related to using multiple item formats and for item technical and editorial quality. However, fewer than half of reviewed items align to at least one of the assigned standards, raising significant concerns around standards alignment to the CCSS.

Item Formats and Technical/Editorial Quality

The ACT English, Reading, and Writing subtests fared well on the use of multiple formats and the editorial quality of the items. Across the subtests, each form uses two item formats: 99 percent of items are traditional multiple-choice items, and the writing prompt in the Writing subtest provides a constructed-response item on each test form that requires students to generate a response. Nearly all items (more than 95 percent on each form) are free from technical quality and editorial accuracy errors. Reviewers noted the exceptional editorial accuracy and technical quality of the items.

Alignment

Overall, reviewers determined that 40 percent of items align to at least one assigned standard. As ACT aligns each English, Reading, and Writing item to multiple standards, reviewers also considered whether the item aligns to each targeted standard and agreed overall with 21 percent of the item-to-standard alignment provided for Form 1 and 24 percent of item-to-standard alignments provided for Form 2.

Table 7. Alignment Index: Agreement with ACT-Claimed Alignment and Percentage of Items That Align to At Least One ACT-Claimed Standard

	% agree with ACT-claimed alignment	% items that align to at least one ACT-claimed standard
Overall	23%	40%
Form 1	21%	50%
Form 2	24%	29%

Reviewers also considered the alignment for each ACT subtest to describe how the alignment of items to standards across the Reading, Writing, and English subtests contributed to the overall alignment index. Table 8 below shows alignment indices for the Reading and English subtests, including both the percentage of reviewer agreement with ACT-claimed alignment and the percentage of items that align to at least one of ACT's claimed standards.

Table 8. Alignment Index: Agreement with ACT-Claimed Alignment and Percentage of Items That Align to At Least One ACT-Claimed Standard on Reading and English Subtests

	% agree with ACT-claimed alignment	% items that align to at least one ACT-claimed standard
Reading subtest		
Overall	4%	12%
Form 1	4%	13%
Form 2	3%	10%
English subtest		
Overall	28%	66%
Form 1	30%	71%
Form 2	26%	60%

Reviewers also considered the single writing prompt included on each test form. Reviewers found that ACT Writing items are aligned to both primary Writing standards (e.g., W.11-12.1) and associated substandards (e.g., W.11-12.1c) that are intended to be contextualized by the primary standards. Because ACT explicitly aligned its writing prompts to both primary standards and substandards, reviewers rated their alignment independently and found that while the primary standard is not assessed, the prompt does assess some of the substandards, as reflected in Table 9 below. It should be noted that the lack of alignment to the primary standard substantially weakens the alignment of the prompts to standards overall.

Table 9. Alignment Index: Agreement with ACT-Claimed Alignment and Whether Prompts Align to At Least One Writing Standard on Writing Subtest[10]

	How many of the ACT-claimed alignments did the reviewers agree with?	Do the prompts align to at least one Writing standard?
Overall	3 out of 7 total standards	Yes
Writing prompt, Form 1	3 out of 7 total standards	Yes
Writing prompt, Form 2	3 out of 7 total standards	Yes

[10] Standards reviewed included the primary and associated substandards as separate standards.

Mathematics: Content Criteria

Achieve's analysis found that the ACT Math subtest does not strongly emphasize content in mathematics as defined by the expectations in the CCSS. Summary ratings are presented in the chart below, followed by a discussion of the components that make up the Content rating (C.1 and C.2). The ACT balances score points across conceptual, procedural, and application items but does not focus on the content most important for high school as described by the Widely Applicable Prerequisites. Fewer than a third of the items on the ACT Math subtest align to the Widely Applicable Prerequisites. While the ACT was found to appropriately balance items across conceptual, procedural, and application tasks, C.1 is much more heavily weighted in the overall rating, resulting in a Weak Match for content.

Mathematics: Overall Content Rating and Criterion Ratings

Criterion	Rating
C.1: Focusing strongly on the content most needed for success in later mathematics Fewer than 30 percent of items on the ACT Math subtest align to the Widely Applicable Prerequisites. The documentation should indicate alignment to the mathematics content progressions documents but does not.	**Weak Match**
C.2: Assessing a balance of concepts, procedures, and applications The ACT balances score points across conceptual, procedural, and application tasks. Documentation lacks clarity on balancing concepts, procedures, and applications.	**Excellent Match**

Criterion C.1: Focusing Strongly on the Content Most Needed for Success in Later Mathematics

Assessments help educators keep students on track to readiness by focusing strongly on the content most needed in each grade or course for later mathematics. As such, the vast majority of score points in an assessment should focus on the content that is most important for students to master in that grade band to reach college and career readiness. For a high school-level assessment, the expectation in this criterion is that at least 50 percent of the score points align to the content outlined in the Widely Applicable Prerequisites found in the *High School Publishers' Criteria for the Common Core State Standards for Mathematics* (National Governors Association et al, 2013). Additionally, there should be sufficient documentation that reflects adherence to the mathematics in the CCSS mathematics content progressions documents (Common Core Standards Writing Team, 2013). For the outcomes review, the reviewers scored items as meeting the Widely Applicable Prerequisites if the alignment, as determined by the reviewers, (a) directly matched the prerequisites, (b) were mentioned as relatively important to the prerequisites (and were not [+] standards), or (c) were directly connected to key takeaways from grades 6–8 at a level of sophistication appropriate to high school.

Overall, only 27 percent of the ACT math items reviewed across two forms meet the expectations of the Widely Applicable Prerequisites (32 percent of the items on Form 1 and 22 percent of the items on Form 2). Additionally, the 2014 *ACT Technical Manual* provides standards with score progressions; however, the reviewers found no clear alignment to the Widely Applicable Prerequisites or the CCSS mathematics content progressions documents. Finally, as described in the *ACT Technical Manual Supplement*, 40–43 percent of ACT mathematics items are intended to measure content below the 8th grade level (Integrating Essential Skills), so

the design of the ACT inherently limits the number of items intended to measure high school mathematics content.

As described in the *ACT Technical Manual Supplement*, ACT has adopted new reporting categories for the Preparing for Higher Math section of the test that appear to align to the conceptual categories (the highest level in the CCSS organization of standards). However, the additional detail is insufficient to provide evidence of how score points may directly align to the Widely Applicable Prerequisites.

Criterion C.2: Assessing a Balance of Concepts, Procedures, and Applications

An assessment should aim to test a balance of concepts, procedures, and applications. In this study, reviewers categorized individual items as targeting one or more of these aspects. To fully meet this criterion, items within each form need to demonstrate the following balance of concepts, procedures, and applications:

- 15–35 percent of the score points are specific to procedural skills alone;
- 25–50 percent of the score points are from conceptual understanding items or items that target both concepts and procedures; and
- 25–50 percent of the score points are from application items that may be connected to procedural skills, conceptual understandings, or both.

The ACT is an excellent match for criteria assessing the balance of concepts, procedures, and applications. Both reviewed forms fall within the ranges above, reflecting a balance of concepts, procedures, and applications that fully meets the criterion. In the documentation reviewers did note a lack of clarity on balancing concepts, procedures, and applications. For example, the *ACT Technical Manual*, published in 2014, makes mention of "Cognitive Levels," which sound similar to the aspects in this category,[11] but the *ACT Technical Manual Supplement*, published in 2016, does not show those levels as part of the current math test blueprints.

Mathematics Depth Criteria

Achieve's analysis finds that ACT Math subtests do not strongly emphasize the depth of the expectations in the CCSS. Summary ratings are presented in the chart below, followed by a discussion of the components that make up the Depth rating (C.3, C.4, and C.5). The items on the ACT Math subtest are largely well written, with relatively few instances of editorial issues. The range of item types included on the assessment is limited, with the full assessment consisting of multiple-choice items. Reviewers agreed with fewer than 50 percent of ACT's claimed item alignments to mathematical content standards. While an appropriate proportion of the assessment reflects mathematical practices, many of those items do not connect to mathematical content standards. Together, these findings resulted in a Weak Match overall for Depth.

[11] Knowledge and skills, direct application, understanding concepts, and integrating understanding (p. 6)

Mathematics: Overall Depth Rating and Criterion Rating

Criterion	Rating
C.3: Connecting practice to content Some, but not the large majority, of items that reflect a mathematical practice also align to mathematical content. Additionally, the provided documentation shows a limited focus on the Standards for Mathematical Practice, as the documentation refers only to modeling and does not discuss the other practices as targets for the blueprint or item design.	**Weak Match**
C.4: Requiring a range of cognitive demand Reviewers agreed with approximately half of the DOK levels claimed by ACT.	**[Not Rated]**
C.5: Ensuring high-quality items and a variety of item types While most items are of high quality (more than 90 percent), some items have editorial errors due to readability and the mathematics. The ACT Math subtest uses traditional multiple-choice items. Additionally, a limited number of items (fewer than 50 percent) are aligned to the claimed CCSS content standard.	**Weak Match**

Criterion C.3: Connecting Practice to Content

The eight Standards for Mathematical Practice are an integral part of the CCSS. Mathematical proficiency for the CCSS involves both content and practices. An assessment that addresses one without the other is missing a large component of the standards. The Standards for Mathematical Practice are listed below:

1. Make sense of problems and persevere in solving them;
2. Reason abstractly and quantitatively;
3. Construct viable arguments and critique the reasoning of others;
4. Model with mathematics;
5. Use appropriate tools strategically;
6. Attend to precision;
7. Look for and make use of structure; and
8. Look for and express regularity in repeated reasoning.

For this criterion, reviewers determined whether each item reflects a practice. Additionally, the analysis included whether items that reflect practices also align to content standards. The goal was that at least 33 percent of the items reflect a practice and that at least 95 percent of those items also align to one or more content standards. For Form 1, 37 percent of the items reflect a practice, but only 64 percent of those also align to a content standard. For Form 2, 33 percent of the items reflect a practice, but only 80 percent of those also align to a content standard.

Table 10. Criterion B.3: Percentage of Items Reflecting a Practice and Percentage of Items Reflecting a Practice that Also Align to a CCSS Content Standard

	% items that reflect a practice	Of those that reflect a practice, % items that also align to a CCSS content standard
Overall	35%	72%
Form 1	37%	64%
Form 2	33%	80%

The set of mathematical practices were not found by reviewers in ACT's provided documentation. In the *ACT Technical Manual Supplement*, a Preparing for Higher Math reporting category for modeling "represents all questions that involve producing, interpreting, understanding, evaluating, and improving models" (p. 1.8). Beyond this emphasis on mathematical modeling, there is no indication of emphasis on the other practices in the test blueprint. Some documentation suggested item-level categories similar to the practices but not how those categories support the development of items or test forms.

Criterion C.4: Requiring a Range of Cognitive Demand

Assessments should require all students to demonstrate a range of higher-order, analytical thinking skills based on the depth and complexity of college- and career-ready standards. A common approach to measuring the cognitive demand of individual assessment items or tasks is the use of Webb's DOK. As developed by Norman Webb (2007), there are four DOK levels, which are briefly described below for mathematics.

- **DOK 1 (Recall):** Includes the recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level.

- **DOK 2 (Skill/concept):** Includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity. These actions imply more than one step.

- **DOK 3 (Strategic thinking):** Requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract.

- **DOK 4 (Extended thinking):** Requires complex reasoning, planning, developing, and thinking, most likely over an extended period of time. At Level 4, the cognitive demands of the task should be high, and the work should be very complex.

Achieve did not determine a rating for C.4 because the methodology did not assign DOK ratings to individual content standards. However, Achieve did calculate the percentage of reviewer agreement with ACT DOK claims for each item and compared these findings with ACT's claimed distribution of DOK in its items. According to the 2016 *ACT Technical Manual Supplement* 12–15 percent of the items should be at DOK 1, 53–60 percent of the items at DOK 2, and 26–34 percent of the items at DOK 3.[12] Achieve reviewers found significantly more items at the DOK 1 level than claimed by ACT and significantly fewer items rated as DOK 3. Additionally, on Form 1, reviewers ranked an item higher than the ACT claim for only two items, while they placed 26 items one DOK level lower. On Form 2, reviewers ranked an item higher than the ACT claim for only three items, while they placed 27 items one DOK level lower. In one instance across both forms, reviewers placed an item two levels below the claim (i.e., the developer-claimed DOK level for an item was 3, and reviewers placed the item at a DOK 1 level).

Table 11. ACT Claims and Achieve Reviewer Ratings by DOK Level for the Mathematics Subtest

	DOK 1	DOK 2	DOK 3
Target as claimed by ACT	12–15%	53–60%	26–34%
Overall (claimed)	15%	61%	24%
Overall (reviewed)	33%	66%	1%
Form 1 (claimed)	12%	63%	25%
Form 1 (reviewed)	28%	70%	2%
Form 2 (claimed)[13]	18%	58%	23%
Form 2 (reviewed)	38%	62%	0%

Criterion C.5: Ensuring High-Quality Items and a Variety of Item Types

Every item on a mathematics assessment should be of high quality; provide evidence for a content standard; and be free of mathematical errors, readability issues, and other problems. Multiple item types should be used. In this criterion at least 90 percent of the items should accurately align to standards, at least 95 percent of the items should be free of editorial and technical flaws, and at least two different item formats should be used. Documentation should support alignment accuracy and quality assurance. ACT's documentation outlines a strong process for assurance of item quality. A similar process likely is still in place even though changes were made to the reporting categories in the 2016 *ACT Technical Manual Supplement*. To be aligned, an item must provide evidence of the student meeting all or part of the indicated standard (or standards) as written. After reading an item and the proposed aligned standard, reviewers determined if the item provides convincing evidence that a student who answers the item correctly has met the outcome indicated by the standard. Items must meet the intention of the standard. It is not sufficient that the

[12] ACT limits the DOK ratings values to 1 to 3.
[13] This line does not sum to 100 percent due to rounding.

mathematics in an item is merely somewhat related to the mathematics in a standard.

Overall, reviewers agreed with fewer than half of ACT's claimed alignments to standards. Across both forms, nine items were found to have quality issues, four due to mathematics and five due to readability. All items are single-select multiple choice and thus do not meet the subcriterion for multiple item formats.

Table 12. Alignment Index: Agreement with ACT-Claimed Alignment to Standards on Mathematics Subtest

	% agree with ACT-claimed alignment
Overall	45%
Form 1	43%
Form 2	47%

RECOMMENDATIONS FOR STATES

The following recommendations for states that would like to use the ACT as their statewide summative assessment are based on the findings of this study. They are intended to support the improvement of a state's high school assessment program so that it more fully reflects the state's adopted academic standards in ELA and mathematics. States should discuss potential improvements with key stakeholders, including the test developer (ACT), education leaders, educators, parents, students, and institutions of higher education. These improvements to the high school assessment program may include changes to the ACT plus Writing as administered to high school students, enhancements to the overall high school assessment program, or both.

Policy Recommendations

States should not use the ACT as the statewide accountability measure for ELA and mathematics. States are using college admissions tests such as the ACT in school accountability systems in multiple ways. Some states are using the test as part of a college readiness indicator, where it is often one of multiple measures that determine student readiness for college; some use it as an equity measure to increase access to college for all students. These uses of college admissions tests in accountability remain appropriate. However, using the ACT as the primary measure of math and ELA achievement of the state college and career readiness standards for accountability is ill advised for both alignment considerations and technical reasons.

States should not allow districts to administer the ACT in lieu of the statewide summative assessment. While the Every Student Succeeds Act gives states the opportunity to provide districts the choice to administer a college admissions test, including the ACT, instead of the statewide summative assessment, states should be cautious about opening this door. Beyond the alignment issues described in this study, Florida's commissioned study of the ACT and SAT (Assessment Solutions Group, 2018) was clear that ensuring comparability between state assessments and college admissions tests will be extremely challenging, if not impossible, for purposes of accountability. With this choice, districts might "shop around" for the assessment that shows them in the best light, while avoiding tough but necessary conversations about the performance of all students.

States that have adopted the ACT should ask ACT to augment their tests to improve alignment to send educators better signals that will influence instruction in a positive manner. This study recommends that the ACT be augmented with additional items that are designed to directly assess states' adopted ELA and mathematics standards to improve alignment. While augmentation presents challenges, particularly added cost and complexity, Achieve urges ACT to respond positively to this recommendation. Augmentation will send more consistent signals to educators about the importance of the content in state-adopted standards, and it will increase the likelihood of the ACT meeting peer review guidelines.

Recommendations for ELA

B.1: **The assessment should include texts that reflect a balance of literary and informational text types.** The CCSS require a 50/50 balance of literary and informational passages by grade, largely based on the 2009 National Assessment of Educational Progress (NAEP) Reading Framework. As such, the passages on the ACT should reflect a more balanced distribution of both literature (stories, drama, and poetry) and informational text, including literary nonfiction.

B.2: **ACT should publish text complexity information.** Assessing students on texts of the appropriate complexity for the grade is the strongest way to determine if students are prepared to meet postsecondary reading demands and is a priority in the CCSS. While the ACT assesses students on grade-appropriate complex text, educators would benefit greatly from understanding how ACT selects texts to inform their instructional practices. Educators need tools and examples of texts that are appropriate to the grade. Thus, ACT should publish sample quantitative and qualitative analyses of texts, so educators can have clear examples of texts that not only reflect the standards but are also similar to texts students will encounter on the assessment.

B.3 **The assessment should require close reading and the use of evidence.** The CCSS directly require that students "read closely to determine what the text says … [and] cite specific textual evidence when writing or speaking to support conclusions drawn from the text." More items on the ACT should require close reading and the use of evidence drawn directly from text to meet the requirements of the standard.

B.5: **The assessment should require students to draw upon text(s) to craft their writing and vary the types of writing that are assessed.** Writing standards 1 and 2 require students to use relevant and sufficient evidence and convey information through the effective selection, organization, and analysis of content. Therefore, students need to use evidence and analyze content in their writing; that is, students need to encounter rich texts on the assessment that can support writing to sources. Additionally, the CCSS require a distribution of writing purposes (to persuade, to explain, to convey experience) as reflected in the 2011 NAEP Writing Framework. To meet this requirement, the ACT should require students to write in a variety of genres (argument, informative/explanatory, narrative) to meet the standards.

B.9: **ELA items should fully align to the grade-level standards.** Reviewers agreed with fewer than 50 percent of ACT's claimed alignments to the CCSS. To make evidence-based claims that the content standards intended for all students have been met, ACT should ensure that its assessment includes more accurate claims of alignment at the item level.

Recommendations for Mathematics

C.1: **More than 50 percent of mathematics items should align to the Widely Applicable Prerequisites or align to key takeaways from grades 6–8 at a level of sophistication appropriate to high school.** Additionally, assessment documentation should indicate alignment to the progressions documents.

C.3. **The assessment should ensure that the Standards for Mathematical Practice are included in assessment design.** These practices are a component of student proficiency and are included in the CCSS.

C.5: **Math items should fully align to the grade-level standards.** Reviewers found fewer than 50 percent of claimed alignments to the CCSS. To make evidence-based claims that the content standards intended for all students have been met, ACT should ensure that its assessment includes more accurate claims of alignment at the item level.

REFERENCES

ACT. (2010a). *ACT technical manual*. Accessed July 12, 2017: https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

ACT. (2010b). *ACT supports Common Core State Standards*. Accessed February 12, 2018: http://www.corestandards.org/assets/ccsi_statements/StatementACT.pdf

ACT. (2015). *How ACT assessments align with state college and career readiness standards*. Accessed July 12, 2017: http://www.act.org/content/dam/act/unsecured/documents/Alignment-White-Paper.pdf

ACT. (2016). *ACT technical manual supplement*. Accessed July 12, 2017: https://www.act.org/content/dam/act/unsecured/documents/ACT-Technical-Manual-Supplement.pdf

Assessment Solutions Group. (2018). *Feasibility of the Use of ACT and SAT in Lieu of Florida Statewide Assessments*. Accessed February 12, 2018: http://www.fldoe.org/core/fileparse.php/5663/urlt/FeasIBILITYactsat.pdf

California State University Early Assessment Program. (2016). Accessed February 12, 2018: https://www.calstate.edu/eap/about.shtml

Common Core Standards Writing Team. (2013). *Progressions documents for the Common Core math standards*. Accessed July 12, 2017: http://ime.math.arizona.edu/progressions/

Common Core State Standards Initiative. *English language arts/literacy and mathematics standards*. Accessed July 12, 2017: http://www.corestandards.org/read-the-standards/

Common Core State Standards Initiative. (2013). *High school publishers' criteria for the Common Core State Standards for mathematics*. Accessed July 12, 2017: http://www.corestandards.org/assets/Math_Publishers_Criteria_HS_Spring%202013_FINAL.pdf

Council of Chief State School Officers. (2014). *Criteria for procuring and evaluating high quality assessments*. Accessed July 12, 2017: http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Council of Chief State School Officers. (2013). *High-quality summative assessment principles for ELA/literacy and mathematics assessments aligned to college and career readiness standards*. Accessed February 12, 2018: https://www.ccsso.org/sites/default/files/2017-12/CCSSO%20Assessment%20Quality%20Principles%2010-1-13%20FINAL.pdf

Darling-Hammond, L., et al. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*.

Thomas B. Fordham Institute. Accessed July 12, 2017: https://edexcellence.net/publications/evaluating-the-content-and-quality-of-next-generation-assessments

National Center for the Improvement of Educational Assessment. (2016). *Guide to evaluating assessments using the CCSSO criteria for high quality assessments: Focus on test content*. Accessed July 12, 2017: http://www.nciea.org/sites/default/files/publications/Guide-to-Evaluating-CCSSO-Criteria-Test-Content_020316.pdf

National Governors Association, Council of Chief State School Officers, Achieve, Council of the Great City Schools, and National Association of State Boards of Education. (2013). *High School Publishers' Criteria for the Common Core State Standards for Mathematics*. Accessed February 21, 2018: http://www.corestandards.org/assets/Math_Publishers_Criteria_HS_Spring%202013_FINAL.pdf

Schultz, S. R., Michaels, H. R., Dvorak, R. N., & Wiley, C. R. H. (2016). *Evaluating the content and quality of next generation high school assessments. Human Resources Research Organization (HumRRO)*. Accessed July 12, 2017: https://www.humrro.org/corpsite/sites/default/files/HQAP_HumRRO_High_School_Study_Final%20Report.pdf

Webb, N. L. (2007). Issues relating to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7–25.

APPENDIX A: ABOUT ACHIEVE

Achieve is an independent, nonpartisan, nonprofit education policy organization dedicated to working with states to raise academic standards and graduation requirements, improve assessments, and strengthen accountability. Created in 1996 by a bipartisan group of governors and business leaders, Achieve is leading the effort to make college and career readiness a priority across the country so that students graduating from high school are academically prepared for postsecondary success. Since 2005, Achieve has been at the forefront of some of the most significant K–12 education standards and assessment projects ever undertaken — working with state teams, governors, state education officials, postsecondary leaders, and business executives to improve postsecondary preparation by aligning key policies with the demands of the real world so that all students graduate from high school with the knowledge and skills they need to fully reach their promise.

Since its founding in 1996, improving the quality and rigor of both state standards and assessments has been a core part of Achieve's mission. Achieve first piloted an assessment review process in 1998, with reviews of Michigan and North Carolina summative tests, and published an overview of its criteria and methodology for evaluating state standards and assessments in 2002 (*Benchmarking and Alignment of State Standards and Testing*). Since then, Achieve has conducted content analyses of assessments for 13 states. These analyses have provided the basis for strategic and technical advice to improve the alignment and rigor of state assessments.

In 2004, Achieve conducted comparative analyses of the content and rigor of high school graduation tests for half a dozen states (*Do Graduation Tests Measure Up? A Closer Look at State High School Exit Exams*) and in 2007 analyzed and reported on the content covered on the SAT; the ACT; Compass; Accuplacer; and 22 additional state-, system-, or institution-level tests used to place students into first-year college courses (*Aligned Expectations? A Closer Look at College Admissions and Placement Tests*).

In addition to reviews and analyses of assessments, Achieve has experience in overseeing the development of multistate assessments aligned to college- and career-ready standards. In 2007 Achieve helped 14 states develop the American Diploma Project Algebra II end-of-course exam and from 2010 to 2014 served as the project management partner for PARCC.

APPENDIX B: METHODOLOGY

The methodology used in this alignment study builds on the methodology developed by the Center for Assessment (2016) and used by Fordham and HumRRO in a previous study (2016). Based on recommendations from this previous work, Achieve modified the methodology slightly (see Appendix C for a discussion of changes to the original methodology).

Both math and ELA reviews involved a multistep review process, consisting of the following stages:

Outcomes review: During the outcomes review phase, reviewers independently considered each item on two test forms for evidence of meeting the appropriate criteria, based on operationalizing the math and ELA alignment criteria described by CCSSO (2014). During these reviews, reviewers both rated individual items based on established parameters (e.g., selecting from a dropdown menu of options) and made descriptive comments to provide a rationale for their ratings.

Form roll-up: Once reviewers had considered individual items, scores assigned to items on each form were aggregated to provide scores at the criterion and subcriterion levels to represent how the test form as a whole matched the expectations of the criterion. Scores were aggregated at the subcriterion level based on the tentative scoring guidance described in the methodology on a scale of 0–2 (corresponding to Does Not Meet, Partially Meets, and Meets the subcriterion), and reviewers considered whether the score matched their individual assessment of the entire test form. If not, the reviewer was able to use his or her professional judgment to provide an alternative subcriterion score and comments describing his or her rationale.

Following individual reviews, reviewers came to consensus at the item, subcriterion, and criterion levels. Once the consensus subcriterion scores were determined, scores were aggregated again to provide a criterion-level score for the test form on the following scale: Excellent, Good, Limited/Uneven, Weak. Reviewers, as a team, were able to review those tentative scores determined by the quantitative judgments they made to determine whether they agreed that the score matched their collective evaluation of both test forms. Finally, these criterion scores across multiple test forms were aggregated to provide Content and Depth scores.

Generalizability review: Reviewers considered the documentation provided by ACT in tandem with their independent outcomes reviews, following a similar process to that detailed above. During this phase, reviewers considered whether the documentation — test blueprints, specifications for item and passage design, quality control processes, etc. — provided sufficient evidence of whether the findings from their item and form review were generalizable to the assessment program as a whole. Reviewers provided descriptive evidence and assigned each subcriterion a score (0–2 or insufficient evidence) based on their findings. The evidence and resulting scores were also discussed as part of developing consensus and are represented as part of the final criterion, Content, and Depth scores.

Criterion-Specific Methodology: ELA

ELA includes nine criteria, each of which are further detailed by 39 total subcriteria. The methodology described here does not address B.8 and its associated subcriteria because they were not evaluated for this study.

Criteria:

- B.1: Using a balance of high-quality literary and informational texts
- B.2: Focusing on the increasing complexity of texts across grades
- B.3: Requiring students to read closely and use evidence from text
- B.4: Requiring a range of cognitive demand
- B.5: Assessing writing
- B.6: Assessing vocabulary and language skills
- B.7: Assessing research and inquiry
- B.9: Ensuring high-quality items and a variety of item types

B.1 and B.2: Passage Quality, Balance, and Complexity

B.1: Using a Balance of High-Quality Literary and Informational Texts

Criterion B.1 consists of six subcriteria that focus on the balance and quality of the texts students read as part of the assessment. The subcriteria relate to the major focus and advances associated with college- and career-ready ELA standards, including the CCSS: the balance of informational and literary texts, the quality of the passages included, and the key features and types of informational texts students read and to which they respond.

Subcriteria:

- B.1.1: Balance of informational and literary passages
- B.1.2: High-quality passages
- B.1.3: Balance of different types of informational passages
- B.1.4: Balance of text types (generalizability)
- B.1.5: Text quality (generalizability)
- B.1.6: Informational text structure and content (generalizability)

Outcomes review: For B.1, reviewers consider all passages associated with reading items. For each passage, reviewers make expert judgments about whether the passage was previously published, is content rich, exhibits exceptional craft, and/or provides useful information and whether the passage is informational (nonfiction) or literary (fiction). For informational texts, reviewers determine whether the passage is primarily narrative or expository and the discipline (science/technology, history/social science, literary nonfiction) with which the text is best associated.

Reviewer decisions about each text are aggregated to a tentative score based on the following scoring guidance for subcriteria on the high school assessments:

Table 1. Scoring Guidance for B.1: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 60–72% of the texts are informational. • 90–100% of the passages are of high quality. • 90–100% of informational texts are expository AND are split nearly evenly across disciplines.	• 40–59% of the texts are informational. • 75–89% of the passages are of high quality. • 75–89% of informational texts are expository AND/ OR address only two of the three disciplines.	• 0–39% of the texts are informational. • 0–74% of the passages are of high quality. • 0–74% of informational texts are expository AND/ OR address only one of the three disciplines.

Generalizability review: Reviewers evaluate the documentation provided for similar criteria, which parallel the outcomes criteria, as described in the scoring guidance below.

Table 2. Scoring Guidance for B.1: Generalizability Review

Meets	Partially Meets	Does Not Meet
• 60–72% of the passages should be informational. • Documentation thoroughly outlines the expectations for and process used to determine that high-quality texts are placed on the assessment. • Documentation outlines the expectation that 90–100% of informational texts are expository AND are split nearly evenly across disciplines.	• 40–59% of the passages should be informational, or 72–90% of the passages should be informational. • Documentation partially outlines the expectations for and process used to determine that high-quality texts are placed on the assessment. • Documentation outlines the expectation that fewer than 90% of informational texts should be expository, OR for grades 6–12, the informational texts are split nearly evenly for the three disciplines mentioned above.	• 0–39% of the passages should be informational, or 91–100% of the passages should be informational. • Documentation does not outline the expectations for and process used to determine that high-quality texts are placed on the assessment. • Documentation does not provide guidance around the structure of informational texts or the distribution of informational text type.

B.2: Focusing on the Increasing Complexity of Texts across Grades

The focus of this criterion is on text complexity: whether the passages on the test forms reviewed are of appropriate complexity for the grade and whether the documentation indicates extensive qualitative and quantitative measures used to determine the appropriate complexity for a given grade.

Subcriteria:

- B.2.1: Passage placed at an appropriate grade band and level
- B.2.2: Appropriately complex texts
- B.2.3: Text complexity process

Outcomes review: For B.2, reviewers evaluate the individual passages found on the reviewed test forms and their associated metadata to determine whether those specific passages have been placed at the appropriate grade band and level, as justified by the text complexity information provided. Reviewers consider both the qualitative and quantitative (e.g., Lexile levels) information provided when making these judgments. Scoring at the form level for this subcriterion follows this guidance:

Table 3. Scoring Guidance for B.2: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 90–100% of passages have been placed at a grade band and level justified by complexity data.	• 75–89% of passages have been placed at a grade band and level justified by complexity data.	• 0–74% of passages have been placed at a grade band and level justified by complexity data.

Generalizability review: For B.2, reviewers determine whether the documentation includes a sufficient process to select and provide quality control for the texts used in the assessment program, using both qualitative and quantitative measures. The scoring guidance is relatively straightforward and parallels the item/outcomes review:

Table 4. Scoring Guidance for B.2: Generalizability Review[14]

Meets	Partially Meets	Does Not Meet
• The documentation thoroughly explains how quantitative data and qualitative analysis are used to ensure that texts are placed at the appropriate grade band and grade level.	• The documentation partially explains how quantitative data and qualitative analysis are used to ensure that texts are placed at the appropriate grade band and grade level.	• The documentation does not explain how quantitative data and qualitative analysis are used to ensure that texts are placed at the appropriate grade band and grade level, or the documentation outlines quantitative ranges that are not supported by research.

[14] Note that the generalizability subcriterion B.2.3, which deals with the process associated with determining the progressions of text complexity over grade levels and grade bands, was not applicable to the ACT because the test program is designed for a single high school grade band (grades 11–12).

B.3–B.9: Item Content, Depth, and Quality

B.3: Requiring Students to Read Closely and Use Evidence from Text

The focus of this criterion is to highlight the important elements of reading as described by college- and career-ready standards like the CCSS. The subcriteria associated with B.3 emphasize the use of direct textual evidence, close reading, and focusing on central ideas.

Subcriteria:

- B.3.1: Require close reading
- B.3.2: Focus on central ideas
- B.3.3: Align to the specifics of the grade-level standards
- B.3.4: Require the use of direct textual evidence
- B.3.5: Documentation indicates that the content of reading items aligns to grade-specific standards (generalizability)
- B.3.6: Documentation indicates that reading items require students to use direct textual evidence (generalizability)

Outcomes review: For B.3, reviewers consider the items coded to reading standards. For each item, reviewers determine whether the item meets the subcriteria by making judgments about:

- Whether an item requires close reading and analysis. Reviewers consider whether the item requires that students carefully read the text and make inferences based on the text to answer the question — in other words, whether the question requires more than simply matching text from the question to texts provided, simple paraphrasing of text, or prior/background knowledge.

- Whether an item requires students to use direct textual evidence. Reviewers consider whether the item requires students to provide or cite quotes or paraphrase directly from the text to support a claim or make an inference about the text.

- Whether an item focuses on central ideas or important particulars of the text, rather than superficial or tangential details.

- Whether an item aligns to the grade-level specifics of the reading standards to which the developer claims they are aligned.

Scores for individual items are aggregated to provide form-level subcriterion scores. To fully meet this criterion, items on each test form are expected to meet the following guidelines:

Table 5. Scoring Guidance for B.3: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 90–100% of items require close reading and analysis of text. • 90–100% of items focus on central ideas and important particulars. • 90–100% of items are aligned to the specifics of the grade-level standards listed. • 51–100% of the reading score points are based on items requiring direct use of textual evidence.	• 75–89% of items require close reading and analysis of text. • 75–89% of items focus on central ideas and important particulars. • 75–89% of items are aligned to the specifics of the grade-level standards listed. • 33–50% of the reading score points are based on items requiring direct use of textual evidence.	• 0–74% of items require close reading and analysis of text. • 0–74% of items focus on central ideas and important particulars. • 0–74% of items are aligned to the specifics of the grade-level standards listed. • 0–32% of the reading score points are based on items requiring direct use of textual evidence.

Generalizability review: Reviewers consider how well the documentation reflects the outcomes subcriteria, including the degree to which it indicates that central ideas, close reading, aligning to grade-level standards, and using textual evidence are required by the assessment program. To meet these subcriteria, the generalizability review is expected to follow this scoring guidance:

Table 6. Scoring Guidance for B.3: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation thoroughly outlines the requirement that reading items arise from close reading and analysis of text, focus on central ideas and important particulars, and assess the depth and specific requirements delineated in the standards at each grade level. • Documentation indicates that 51–100% of the reading score points should be based on items requiring direct use of textual evidence.	• Documentation partially outlines the requirement that reading items arise from close reading and analysis of text, focus on central ideas and important particulars, and assess the depth and specific requirements delineated in the standards at each grade level. • Documentation indicates that 33–50% of the reading score points should be based on items requiring direct use of textual evidence.	• Documentation does not outline the requirement that reading items arise from close reading and analysis of text, focus on central ideas and important particulars, and assess the depth and specific requirements delineated in the standards at each grade level. • Documentation indicates that 0–32% of the reading score points should be based on items requiring direct use of textual evidence.

B.4: Requiring a Range of Cognitive Demand

The focus of this criterion is to determine whether the assessment items on a given test form, as well as the assessment program as a whole, reflect a range of cognitive demand that is sufficient to assess the depth and complexity of the state's standards, as evidenced by use of a generic taxonomy (e.g., Webb's DOK) or, preferably, classifications specific to the discipline and drawn from mathematical factors. Because there is no clear target distribution of cognitive demand that assessments should necessarily hit — and therefore no way to justifiably determine whether a given distribution fully or partially meets this criterion — B.4 is not assigned an overall subcriterion or criterion-level score. Instead, reviewers consider to what extent the DOK of the items match the developer-claimed DOK and whether the documentation indicates processes to determine DOK and ensure that a range is represented on all forms across the assessment program.

Subcriteria:

- B.4.1: Determining the distribution of cognitive demand
- B.4.2: Determining the distribution of cognitive demand (generalizability)

Outcomes review: Reviewers use the DOK rating scale to independently assign each item a DOK level, without knowledge of the developer's claimed DOK assignment. After reaching consensus at the item level, reviewer-determined item-level DOKs are compared to the developer's claimed DOK to describe the degree to which the independent evaluation of item-level DOK matches the developer's claims about DOK.

Generalizability review: Reviewers consider the documentation to determine the extent to which assessments are designed to include a distribution of cognitive demand such that for each grade level, the range is sufficient to assess the depth and complexity of the state's standards (e.g., complexity initial targets, evaluation process, and processes for editing DOK levels of items as the assessment is developed).

B.5: Assessing Writing

B.5 focuses on writing prompts that require students to engage in close reading and analysis of texts so that students can demonstrate college- and career-ready abilities. The subcriteria that comprise B.5 focus on a balance of writing types across exposition, argument, and narrative types and on the extent to which writing prompts require students to write to sources.

Subcriteria:

- B.5.1: Distribution of writing types
- B.5.2: Text-based writing
- B.5.3: Distribution of writing types (generalizability)
- B.5.4: Writing to sources (generalizability)

Outcomes review: Reviewers consider the writing prompts (i.e., those items that require students to produce writing). They evaluate each writing prompt to determine which type of writing (exposition, argument, narrative) is called for and determine whether the writing prompt requires students to confront text or other relevant stimuli directly, draw on textual evidence, and support valid inferences from text or stimuli. This information is aggregated to the form level to determine how different writing types are distributed on the form as well as the proportion of items that require students to engage in text-based writing. To fully meet these subcriteria, test forms are expected to demonstrate the following:

Table 7. Scoring Guidance for B.3: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• All three writing types (exposition, argument, and narrative) are present, each representing 28–38% of writing prompts. Note that in high school, the balance is expected to shift toward exposition and argument; if narrative writing is not assessed, the assessment can still receive a 2. • 90–100% of writing prompts require students to confront text or other stimuli directly, draw on textual evidence, and support valid inferences from text or stimuli.	• Two of the three writing types are present. • 75–89% of writing prompts require students to confront text or other stimuli directly, draw on textual evidence, and support valid inferences from text or stimuli.	• Only one writing type is present across all forms. • 0–74% of writing prompts require students to confront text or other stimuli directly, draw on textual evidence, and support valid inferences from text or stimuli.

Generalizability review: The generalizability review mirrors the outcomes review as reviewers determine whether the documentation specifies the distribution of various writing types and a balance shifting toward more exposition and argument at the higher grade levels and whether the specifications require students to write to sources.

To fully meet the B.5 generalizability subcriteria, assessment program documentation needs to demonstrate the following:

Table 8. Scoring Guidance for B.5: Generalizability Review

Meets[15]	Partially Meets	Does Not Meet
• Documentation indicates that all three writing types are approximately equally represented in the grade band, allowing blended types to contribute to the distribution. • Documentation indicates that 51–100% of the reading score points should be based on items requiring direct use of textual evidence.	• Documentation indicates that two of the three writing types are represented in the grade band, allowing blended types to contribute to the distribution. • Documentation indicates that 33–50% of the reading score points should be based on items requiring direct use of textual evidence.	• Documentation indicates that one of the three writing types is represented in the grade band. • Documentation indicates that 0–32% of the reading score points should be based on items requiring direct use of textual evidence.

B.6: Emphasizing Vocabulary and Language Skills

This criterion focuses on assessing proficiency in the use of language, including vocabulary and conventions that reflect college and career readiness. The eight subcriteria associated with B.6 accordingly focus on features of the vocabulary and language items and the percentage of score points associated with each on an assessment.

Subcriteria:

- B.6.1: Vocabulary items reflect requirements for college and career readiness
- B.6.2: Language assessments reflect college and career readiness
- B.6.3: Assessments place a sufficient emphasis on vocabulary
- B.6.4: Assessments place a sufficient emphasis on language
- B.6.5: Vocabulary items reflect requirements for college and career readiness (generalizability)
- B.6.6: Language assessments reflect college and career readiness (generalizability)
- B.6.7: Assessments place a sufficient emphasis on vocabulary (generalizability)
- B.6.8: Assessments place a sufficient emphasis on language (generalizability)

[15] For high school (grades 9–12) programs that do NOT include narrative writing:

2 – Meets: Documentation indicates that expository and argument writing types should be approximately equally represented in the grade band, allowing blended types to contribute to the distribution.

1 – Partially Meets: Documentation indicates that both writing types should be represented but one much more heavily than the other (i.e., one writing type accounts for more than 70 percent of the balance) OR that no balance between the two is outlined.

0 – Does Not Meet: Documentation indicates that only one writing type (expository OR argument) should be represented.

Outcomes review: Reviewers consider both items on test forms and the associated metadata for their review of the B.6 subcriteria. For the vocabulary subcriteria, reviewers consider whether vocabulary items focus on Tier 2 words, ask students to use context to determine the meaning of words, and assess words that are important to the central ideas of the text. Additionally, reviewers determine whether vocabulary is a reporting category and how many score points are devoted to vocabulary on assessment forms.

Similarly, reviewers consider both test forms as well as associated metadata for the language items. Reviewers evaluate whether language items mirror real-world activities (such as actual editing, revision, production of writing), focus on common student errors, and focus on those conventions most important for readiness as described in the CCSS Language Skills Progression Chart. Reviewers also determine whether language is a reporting category for the assessment and how many score points are devoted to language items on assessment forms.

This information is aggregated to produce a form-level score for each subcriterion, which are aggregated to criterion-level scores. To fully meet the criteria by way of the subcriteria, assessment forms are expected to demonstrate the following:

Table 9. Scoring Guidance for B.3: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 75–100% of vocabulary items focus on Tier 2 words AND require use of context, and 51–100% of vocabulary items assess words that are important to central ideas. • 75–100% of the items in the language skills component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions that are most important for readiness. • Vocabulary is reported as a statistically reliable subscore, OR more than 13% of score points are devoted to assessing vocabulary. • Language skills are reported as a statistically reliable subscore, OR more than 13% of score points are devoted to assessing language skills.	• 50–74% of vocabulary items focus on Tier 2 words AND require use of context, and/or 33–50% of vocabulary items assess words that are important to central ideas. • 50–74% of the items in the language skills component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions that are most important for readiness. • 10–12% of score points are devoted to assessing vocabulary. • 10–12% of score points are devoted to assessing language.	• 0–49% of vocabulary items focus on Tier 2 words AND require use of context, and/or 0–32% of vocabulary items assess words that are important to central ideas. • 0–49% of the items in the language skills component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions that are most important for readiness. • 0–9% of score points are devoted to assessing vocabulary. • 0–9% of score points are devoted to assessing language.

Generalizability review: The generalizability review parallels the outcomes review, as reviewers consider whether the test specifications and other documentation specify a focus on each of the features described in the outcomes criteria. To fully meet the documentation subcriteria, assessment program documentation needs to provide evidence of the following:

Table 10. Scoring Guidance for B.6: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation thoroughly outlines the requirement that vocabulary items focus on general academic (Tier 2) words, ask students to use context to determine meaning, and assess words that are important to the central ideas of texts. • Documentation thoroughly outlines the requirement that language is assessed in the writing rubric or that language skills items mirror real-world activities and focus on common student errors and the conventions that are most important for readiness. • Documentation indicates that vocabulary is reported as a statistically reliable subscore or more than 13% of score points. • Documentation indicates that language skills are reported as a statistically reliable subscore OR more than 13% of score points.	• Documentation partially outlines the requirement that vocabulary items focus on general academic (Tier 2) words, ask students to use context to determine meaning, and assess words that are important to the central ideas of texts. • Documentation partially outlines the requirement that language is assessed in the writing rubric or that language skills items mirror real-world activities and focus on common student errors and the conventions that are most important for readiness. • Documentation indicates that vocabulary items will comprise of 10–12% of score points. • Documentation indicates that language skills items will comprise 10–12% of score points.	• Documentation does not outline the requirement that vocabulary items focus on general academic (Tier 2) words, ask students to use context to determine meaning, and assess words that are important to the central ideas of texts. • Documentation does not outline the requirement that language is assessed in the writing rubric or that language skills items mirror real-world activities and focus on common student errors and the conventions that are most important for readiness. • Documentation indicates that vocabulary items will comprise of 0–9% of score points. • Documentation indicates that language skills items will comprise less than 10% of score points.

B.7: Assessing Research and Inquiry

This criterion focuses on whether assessments ask students to demonstrate research and inquiry skills, as demonstrated by the ability to find, process, synthesize, organize, and use information from sources.

Subcriteria:

- B.7.1: Assessing research and inquiry
- B.7.2: Emphasis on research (generalizability)
- B.7.3: Content of research tasks (generalizability)

Outcomes review: Research and inquiry items mirror real-world activities and require students to analyze, synthesize, organize, and use information from sources, such as those tasks that require writing to sources and selecting, analyzing, and organizing evidence from more than one source (and often from sources in different formats). Reviewers consider both test forms and the associated metadata to determine whether items meet this criterion. Additionally, they consider the total number of research items and score points and, of those items and score points, the proportion that mirror real-world activities.

Reviewers' evaluations of items for this subcriterion are aggregated at the form level and subsequently across forms to the criterion level. Subcriterion scoring follows this guidance:

Table 11. Scoring Guidance for B.7: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
75–100% of research prompts require analysis, synthesis, and/or organization of information.	51–74% of research prompts require analysis, synthesis, and/or organization of information.	0–50% of research prompts require analysis, synthesis, and/or organization of information. NOTE: If there is no research component, score this as 0.

Generalizability review: Reviewers consider test blueprints and other specifications as well as exemplar test items for each grade level provided, demonstrating that when assessment constraints permit, real or simulated research tasks comprise a significant percentage of score points when all forms of the Reading and Writing test are considered together. Reviewers look for evidence of an emphasis on research as well as for specifications about the content of research tasks. The generalizability review follows this scoring guidance:

Table 12. Scoring Guidance for B.7: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation indicates that the program reports a research score or otherwise demonstrates that research is significant. • Documentation thoroughly outlines the requirement that research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source and often from sources in diverse formats.	• Documentation indicates that the program includes research tasks but does not indicate that research is significant. • Documentation partially outlines the requirement that research tasks require writing to sources, including analyzing, selecting, and organizing evidence from more than one source and often from sources in diverse formats, but it does not clearly indicate a process for ensuring that items meet this expectation.	• No research tasks are specified to be included. • Documentation does not outline the requirement that research tasks require writing to sources, including analyzing, synthesizing, and organizing evidence from more than one source.

B.9: Ensuring High-Quality Items and a Variety of Item Types

This criterion focuses on three major aspects of items within an assessment: the technical and editorial quality of items, the variety of item types on forms, and the alignment of items to standards.

Subcriteria:

- B.9.1: Variety of item types
- B.9.2: Technical and editorial quality and alignment to standards
- B.9.3: Distribution of item type (generalizability)
- B.9.4: Item alignment and technical and editorial quality

Outcomes review: Reviewers consider test forms and metadata to ensure that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed.[16] Reviewers consider the item types indicated in the metadata and determine whether there are at least two item types and whether at least one of those item types requires students to generate (rather than select) a response.

Reviewers also consider whether the items are of high technical and editorial quality and whether they align to standards. For each item, reviewers determine whether any editorial issues are present and whether they agree or disagree with the developer's alignment claim for each standard identified by the developer. To be considered aligned to standards overall, reviewers must agree with at least one item-to-standard alignment claimed by the developer.

[16] Item types may include, for example, selected-response, two-part evidence-based selected-response, short and extended constructed-response, technology-enhanced, and performance tasks.

Reviewer scores at the item level are aggregated to provide subcriterion scores (described below) and criterion-level scores.

Table 13. Scoring Guidance for B.9: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., constructed response, extended writing). • 90–100% of the items accurately align to standards, and 95–100% of the items are free of editorial and technical flaws.	• At least two formats are used, but the item formats require students only to select, rather than generate, a response (e.g., technology-based formats, two-part selected-response formats). • 80–89% of the items accurately align to standards, and/or 90–94% of the items are free of editorial and technical flaws.	• Only a traditional multiple-choice format is used. • 0–79% of the items accurately align to standards, and/or 0–89% of the items are free of editorial and technical flaws.

Generalizability review: Reviewers conduct a parallel review of the documentation to determine whether the provided specifications detail a distribution of varied item types and provide support for claims about item alignment and the technical and editorial quality of items. The generalizability review follows this scoring guidance:

Table 14. Scoring Guidance for B.9: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation indicates that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., constructed response, extended writing). • Documentation thoroughly supports claims that items are accurately aligned to standard and do not have technical or editorial flaws.	• Documentation indicates that at least two formats are used, but the item formats require students only to select, rather than generate, a response (e.g., technology-based formats, two-part selected-response formats). • Documentation partially supports claims that items are accurately aligned to standard and do not have technical or editorial flaws.	• Documentation indicates that only a single format should be used, including the traditional multiple-choice format. • Documentation does not support claims that items are accurately aligned to standard and do not have technical or editorial flaws.

Mathematics

There are five criteria in mathematics, each of which are further detailed by seven total subcriteria.

Criteria:

- C.1: Focusing strongly on the content most needed for success in later mathematics
- C.2: Assessing a balance of concepts, procedures, and applications
- C.3: Connecting practice to content
- C.4: Requiring a range of cognitive demand
- C.5: Ensuring high-quality items and a variety of item types

C.1: Focusing Strongly on the Content Most Needed for Success in Later Mathematics

C.1 focuses on college and career readiness by requiring that the vast majority of the items and score points on an assessment focus on the content most needed for later success in mathematics. In high school specifically, this focus means that at least half of the points in each grade/course align exclusively to the prerequisites for careers and a wide range of postsecondary studies, as described in the Widely Applicable Prerequisites document.

Subcriteria:

- C.1.1: Assessing the most important content for readiness
- C.1.2: Assessment design reflects the most important content (generalizability)

Outcomes review: Reviewers consider test forms and the accompanying metadata to determine whether each item on a test form is accurately aligned to a standard and whether the item aligns to the major work of the grade and, in high school, the Widely Applicable Prerequisites. If reviewers disagree with the claimed alignment, they can propose an alternative alignment, if appropriate. Reviewers' judgments about alignment determine the percentage of score points that assess the most important content. Scoring for this subcriterion follows this guidance:

Table 15. Scoring Guidance for C.1: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 50–100% of the score points align exclusively to the Widely Applicable Prerequisites.	• 40–50% of the score points align exclusively to the Widely Applicable Prerequisites.	• 0–39% of the score points align exclusively to the Widely Applicable Prerequisites.

Generalizability review: Reviewers review test documentation, including blueprints and other specifications, to determine whether the assessment design (1) indicates that the vast majority of score points in each assessment focus on the most important content and (2) reflect state standards and a coherent progression of mathematics content from grade to grade and course to course.[17] Reviewers assign a score based on their evaluation of the program documentation according to the following guidelines:

Table 16. Scoring Guidance for C.1: Generalizability Review

Meets[18]	Partially Meets	Does Not Meet
• The test blueprints or other documents indicate that at least half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies. • The test blueprints or other documents thoroughly outline the process used for reflecting adherence to the progressions documents.	• The test blueprints or other documents indicate that nearly half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies. • The test blueprints or other documents partially outline the process used for reflecting adherence to the progressions documents.	• The test blueprints or other documents indicate that fewer than half of score points in each course or grade align exclusively to prerequisites for careers and a wide range of postsecondary studies. • The test blueprints or other documents do not outline the process used for reflecting adherence to the progressions documents.

C.2: Assessing a Balance of Concepts, Procedures, and Applications

This criterion focuses on determining whether an assessment measures conceptual understanding, fluency and procedural skills, and application of mathematics as set out in college- and career-ready standards.

Subcriteria:

- C.2.1: Balance of the percentage of points across conceptual understanding, procedural skills and fluency, and applications
- C.2.2: Balance of the percentage of points across conceptual understanding, procedural skills and fluency, and applications (generalizability)

Outcomes review: Reviewers consider the test forms and metadata to determine what knowledge and skills the items assess and how score points are balanced across the categories. Reviewers make judgments about whether each item assesses one of the following targets:

[17] Progressions documents can be found at ime.math.arizona.edu/progressions.

[18] For high school assessments, items that align to the grade 6–8 prerequisites for careers and a wide range of postsecondary studies meet this criterion as long as they are at a level of sophistication appropriate for high school.

Table 17. Aspects of Rigor

Aspects of Rigor Matrix	The item does not involve application.[19]	The item involves a contrived application.[20] The situation is superficial and is not likely a legitimate application of math in life or the workplace.	The item involves an application that is authentic and illustrative of how mathematics is used in real life.[21] The application drives the mathematics.
The item targets procedural skill expected by the grade level.	**PSF**[22]	**PSF-APPC**[23]	**PSF-APPA**[24]
The item targets conceptual understanding and procedural skill expected by the grade level OR targets conceptual understanding but can also be answered using at least some procedural skill expected by the grade level.[25]	**C-PSF**	**C-PSF-APPC**	**C-PSF-APPA**
The item targets conceptual understanding. Students may explain, strategize, evaluate, determine, compare, or classify.	**C**[26]	**C-APPC**	**C-APPA**

Based on reviewers' assignment of items to one of these categories, the balance of total score points associated with procedural skills/fluency (i.e., total percentage of points for PSF), conceptual understanding (i.e., combined total number and percentage of points for C-PSF and C), and application (combined total number and percentage of points for PSF-APPC, C-PSF-APPC, C-APPC, PSF-APPA, C-PSF-APPA, and C-APPA) is determined. Scoring on this subcriterion follows these guidelines: (See Table 18 on next page.)

[19] Names or mathematical referents (e.g., units of measure) may be present in the item but should not be considered "application" in the sense intended by the shifts in college- and career-ready standards.
[20] "Contrived" means that the situation is forced to accommodate the desired mathematics and is not really a plausible application of the mathematics.
[21] "Authentic" includes items in which a simple change of context would make the mathematics broadly applicable to scenarios encountered in life (i.e., depersonalize the analysis to ensure that the trademark of authenticity is about "life" and not "my life").
[22] PSF: Procedural Skills and Fluency.
[23] APPC: Contrived Application
[24] APPA: Authentic Application
[25] Conceptual understanding refers to the mathematics used to respond to an item, not the complexity of the question itself.
[26] C: Conceptual Understanding.

Table 18. Scoring Guidance for C.2: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• 15–35% of the score points are in PSF-TOTAL, 25–50% of score points are in C-TOTAL, and 25–50% of score points are in APP-TOTAL.	• All three categories (PSF-TOTAL, C-TOTAL, and/or APP-TOTAL) are within 5 percentage points of the lower and upper bounds for the threshold defined in Meets.	• One or more of the calculated percentages for the categories (PSF-TOTAL, C-TOTAL, and APP-TOTAL) falls outside of the ranges defined by score points 2 and 1.

Generalizability review: The generalizability review parallels the outcomes review. Reviewers consider provided test documentation to determine the extent to which the test blueprints or other documentation reflect a balance of mathematical concepts, procedures/fluency, and applications, as the standards require. Reviewers provide a score based on the evidence in the provided documentation according to the following criteria:

Table 19. Scoring Guidance for C.2: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation thoroughly outlines the process used for reflecting a balanced assessment for EACH of the three categories.	• Documentation partially outlines the process used for reflecting a balanced assessment for EACH of the three categories.	• Documentation does not outline the process used for reflecting a balanced assessment for each of the three categories or reflects an inadequate balance of score points for one or more of the three categories.

C.3: Connecting Practice to Content

This criterion focuses on the integration of Standards for Mathematical Practice with content standards by requiring that assessments include brief questions and longer questions that connect the most important mathematical content of the grade or course to mathematical practices (e.g., modeling and making mathematical arguments). The goal is for every test item that assesses mathematical practices to also align to one or more content standards (most often within the Major Work of the grade) and for test items through the grades to reflect growing sophistication of mathematical practices with appropriate expectations at each grade level.

Subcriteria:

- C.3.1 Meaningful connections between practice and content
- C.3.2/3.3 Specifications and explanation of the inclusion of mathematical practices

Outcomes review: Reviewers consider test forms and metadata to determine whether assessments meaningfully connect mathematical practices and processes with mathematical content (especially with the most important mathematical content at each grade). For each item, reviewers determine if the item reflects one or more of the "indicators" found in the Assessing the Standards for Mathematical Practice guidance document. If a mathematical practice is reflected, reviewers determine if it also aligns to a content standard. Based on these judgments, reviewers provide a score based on the following guidance:

Table 20. Scoring Guidance for C.3: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• At least 33% of the items reflect the Standards for Mathematical Practice, and of the items that do, 95–100% also align to one or more content standard(s).	• At least 20–32% of the items reflect the Standards for Mathematical Practice, and of the items that do, 90–95% also align to one or more content standard(s).	• Fewer than 20% of the items reflect the Standards for Mathematical Practice, OR fewer than 90% of the items that reflect the Standards for Mathematical Practice align to content standards.

Generalizability review: The generalizability review parallels the outcomes review. Reviewers consider provided test documentation, such as task templates, scoring templates, and other item specifications and explanatory materials, to determine whether they specify how mathematical practices will be assessed. Reviewers determine a final generalizability score for this criterion based on the following scoring guidance:

Table 21. Scoring Guidance for C.3: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation indicates that the Standards for Mathematical Practice are an important feature of the assessment design. It also indicates that all or nearly all items that assess mathematical practices also align to one or more content standards AND concrete examples are given to illustrate how the Standards for Mathematical Practice are reflected in the items.	• Documentation indicates that the Standards for Mathematical Practice reflect a small portion of the assessment design. It also indicates that all or nearly all items that assess mathematical practices also align to one or more content standards AND/OR concrete examples are given to illustrate how the Standards for Mathematical Practice are reflected in the items.	• Documentation indicates that the Standards for Mathematical Practice are not an important consideration in assessment design. It also pays little attention to the need to connect practice to one or more content standards, AND there are limited or no concrete examples given to illustrate how the Standards for Mathematical Practice are reflected in the items.

C.4: Requiring a Range of Cognitive Demand

This criterion focuses on ensuring that assessments require all students to demonstrate a range of higher-order, analytical thinking skills in math based on the depth and complexity of college- and career-ready standards.

Subcriteria:

- C.4.1: Distribution of cognitive demand
- C.4.2: Distribution of cognitive demand (generalizability)

Outcomes review: Reviewers use the DOK rating scale to independently assign each item a DOK level, without knowledge of the test developer's claimed DOK assignment. After reaching consensus at the item level, reviewer-determined item-level DOKs are compared to the developer's claimed DOK to describe the degree to which the independent evaluation of item-level DOK matches the developer's claims about DOK.

Generalizability review: Reviewers review the documentation to determine the extent to which assessments are designed to include a distribution of cognitive demand such that for each grade level, the range is sufficient to assess the depth and complexity of the state's standards (e.g., complexity initial targets, evaluation process, and processes for editing DOK levels of items as the assessment is developed).

C.5: Ensuring High-Quality Items and a Variety of Item Types

This criterion focuses on ensuring that high-quality items and a variety of item types are strategically used to appropriately assess the standard(s). This criterion focuses on three major aspects of items within an assessment: the technical and editorial quality of items, the variety of item types on forms, and the alignment of items to standards.

Subcriteria:

- C.5.1: Variety of item types
- C.5.2: Technical and editorial quality and alignment to standards
- C.5.3: Distribution of item type (generalizability)
- C.5.4: Item alignment and technical and editorial quality

Outcomes review: Reviewers consider test forms and metadata to ensure that the distribution of item types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed [27] Reviewers consider the item types indicated in the metadata and determine whether there are at least two item types and whether at least one of those item types requires students to generate (rather than select) a response.

Reviewers also consider whether the items are of high technical and editorial quality and whether they align to standards. For each item, reviewers determine whether any editorial issues are present and

[27] Item types may include selected-response, short and extended constructed-response, technology-enhanced, and multistep problems.

whether they agree or disagree with the developer's alignment claim for each standard identified by the developer. To be considered aligned to standards overall, reviewers must agree with at least one item-to-standard alignment claimed by the developer.

Reviewer scores at the item level are aggregated to provide subcriterion scores (described below) and criterion-level scores.

Table 22. Scoring Guidance for C.5: Item/Outcomes Review

Meets	Partially Meets	Does Not Meet
• At least two item formats are used, including one that requires students to generate, rather than select, a response (i.e., constructed response, gridded response). • 90–100% of the items accurately align to standards, and 95–100% of the items are free of editorial and technical flaws.	• At least two formats are used, but the item formats require students only to select, rather than generate, a response (e.g., multiple choice, multiselect). • 80–89% of the items accurately align to standards, and/or 90–94% of the items are free of editorial and technical flaws.	• Only a traditional multiple-choice format is used. • 0–79% of the items accurately align to standards, and/or 0–89% of the items are free of editorial and technical flaws.

Generalizability review: Reviewers conduct a parallel review of the documentation to determine whether the provided specifications detail a distribution of varied item types and provide support for claims about item alignment and the technical and editorial quality of items. The generalizability review follows this scoring guidance:

Table 23. Scoring Guidance for C.5: Generalizability Review

Meets	Partially Meets	Does Not Meet
• Documentation indicates that at least two item formats should be used, including one that requires students to generate, rather than select, a response (i.e., constructed response, gridded response). • Documentation thoroughly supports claims that items are accurately aligned to standard and do not have technical or editorial flaws.	• Documentation indicates that at least two formats are used, but the item formats require students only to select, rather than generate, a response (e.g., multiple choice, multiselect). • Documentation partially supports claims that items are accurately aligned to standard and do not have technical or editorial flaws.	• Documentation indicates that only a single format should be used, including the traditional multiple-choice format. • Documentation does not support claims that items are accurately aligned to standard and do not have technical or editorial flaws.

APPENDIX C: CHANGES TO CENTER FOR ASSESSMENT METHODOLOGY

Overview of Methodology Revisions

Over the past year, Achieve has developed an updated version of the methodology from the Center for Assessment's *Guide to Evaluating Assessments Using the CCSSO Criteria for High Quality Assessments: Focus on Test Content*. Using the recommendations from the Fordham and HumRRO reviews as the basis for the revisions, the Achieve team worked with those involved in the development of the original methodology and the Fordham and HumRRO reviews. In addition, the Achieve team considered the resources required for conducting reviews as well as the intended purpose and context of the planned reviews, which are to provide an individual state with a clear picture of the alignment of its assessment program but more importantly to provide specific information and recommendations that can be used to improve the state's assessment program. The resulting methodology maintains the many strengths of the original methodology and, by integrating learnings from the Fordham and HumRRO review experiences, represents a step forward in evaluating assessments based on the CCSSO *Criteria for Procuring and Evaluating High Quality Assessments*.

Below, Achieve has provided a high-level overview of the changes to the methodology and training and review processes. This overview is not intended to itemize every minor change that was made but instead to highlight the more substantive changes to the methodology.

General Issues and Resultant Changes

In a few of the recommendations, Fordham and HumRRO study reviewers cited a desire to comment on specific aspects of the assessments that are more holistic or do not cleanly fit into a rubric but are important to communicate to those interested in understanding more about an assessment and its overall alignment and quality. For this reason, the methodology has been augmented to include specific questions that will not be evaluated in the rubric or score roll-up but instead will be addressed in the narrative of the report. Some examples include:

- Fordham reviewers expressed concern with not having ability to note enthusiasm for good items. Reviewers will now be encouraged to note high-quality items and will be able to comment on them in the narrative commentary of the report (while preserving test security).

- Fordham reviewers expressed concern over the approach for evaluating coverage of major clusters on mathematics assessments, specifically that it was possible for an assessment to meet criterion C.1 while assessing only one cluster of major work. While the CCSSO criteria do not explicitly prescribe specific balances of major work clusters, it is agreed that extreme imbalances would be problematic and should be surfaced through the review process. A narrative commentary prompt allows reviewers to note such an extreme case or simply to comment on notable aspects of coverage.

- Fordham reviewers expressed concern over the approach for evaluating the variety of item types,

specifically that assessments varied quite a bit, and recommended that there be an ability to give credit for assessment programs that have a wider variety of item types. It is agreed that programs can vary significantly in this respect and that the report should give credit for assessment programs that employ more open-ended items as well as a variety of item types. Like Fordham's approach, Achieve's report will include discussion about the variety of item types and proportion of score points.

Cognitive Complexity (B.4, C.4)

The Fordham and HumRRO reviewers raised some concerns regarding the approach for evaluating cognitive complexity. Specifically, there were questions around the targeted DOK distribution and the validity of those targets with respect to the CCSS. There were concerns that the methodology did not capture item difficulty, and there was reviewer disagreement about what to do when the DOK "match" was too high vs. too low. As a result, an overall rating will not be calculated for B.4 and C.4, though the consensus reviewer DOK selection by item will be reported.

Item Quality and Alignment (B.9, C.5)

Issues were noted by the reviewers in the Fordham/HumRRO studies with respect to evaluating item quality and alignment across subject areas. The criteria that call for high-quality items are B.9 and C.5, but reviewers cited issues pertaining to the definition of "high-quality items" as well as logistical issues such as where reviewers were to note issues with alignment to standards during the review. After discussion and weighing of options, it was decided that clarification of the definition of "quality" in the methodology combined with clearer instructions for reviewers was the best solution. As such, the methodology includes the following definition: High-quality items are those that are accurately aligned to standards and do not have technical or editorial flaws. Additionally, narrative prompts have been added to this area to allow reviewers to comment on strong and weak items or additional aspects of quality.

ELA Issues and Resultant Changes

Evaluation of Single Form/Multiple Forms

The ELA reviewers from the Fordham and HumRRO studies noted several issues that were related to the small sample size of passages and writing prompts when reviewing only one form of an assessment. It was determined that reviewing multiple forms to increase the number of passages and writing prompts evaluated for an assessment was the best solution. In the methodology, a note has been added for the specific subcriteria affected by this issue that these subcriteria are best evaluated across multiple forms. Developers submitting for evaluation will be advised that providing multiple forms for those subcriteria will result in better information. When limited passages and writing prompts are available for review, a note has been added to the methodology to invite reviewers to use professional judgment and information from the generalizability review to make a decision. Additionally, the limitation will be noted in the narrative commentary as a consideration when reviewing and drawing conclusions from the results.

Evaluation of Text Quality (B.1)

In the original methodology, reviewers used a proxy for text quality: whether a text was previously published or commissioned. However, reviewers felt that text quality varied, even among previously published texts. To address this feedback, the scoring guidance in the methodology has been revised to match the language of the criteria themselves regarding high-quality texts. The definition of high-quality is no longer exclusively determined by whether or not a work was previously published but is instead evaluated using the elements included in the CCSSO criteria: is content rich, exhibits exceptional craft and thought, and/or provides useful information.

Mathematics Issues and Resultant Changes

Evaluation of Balance of Rigor (C.2)

Perhaps the most problematic issue raised by the mathematics reviewers in the Fordham review was the confusion around evaluation of criterion C.2, balance of rigor. There was such lack of consistency in the findings of individual reviews that the final evaluation was limited to qualitative discussion of the findings. Through a partnership with Student Achievement Partners, a new methodology has been developed to evaluate the balance of aspects of rigor. This methodology allows items to be aligned to any combination of the aspects of rigor and has been tested with small groups over the last year. Achieve feels confident that through this change and new training materials the earlier confusion will be addressed.

Evaluation of Standards for Mathematical Practice (C.3)

The math reviewers from the Fordham and HumRRO studies noted that the evaluation of C.3 in the original methodology resulted in no differentiation between programs. Due to the limited scope of the C.3 review, every assessment reviewed received full credit for this criterion, regardless of the reviewers' observation of how well the assessment addressed the Standards for Mathematical Practice. Additionally, if an assessment neglected the practice standards entirely, it would still receive full credit. In response to this concern, the methodology has been revised to allow reviewers to first identify whether any Standard for Mathematical Practice is evident in an item and then to identify whether items that reflect practice standards also address content standards. An element has been added to the cut-offs that requires that an adequate percentage of items reflect mathematical practices. This solution will avoid the false positive that resulted from the previous study in which an assessment that does not address the Standards for Mathematical Practice could receive full credit for this criterion.

1919 M Street, NW #450 Washington, DC 20036 • @AchieveInc • @OfficialNGSS