A FRAMEWORK TO EVALUATE COGNITIVE COMPLEXITY IN MATHEMATICS ASSESSMENTS

Background

In 2013, the Council of Chief State School Officers (CCSSO), working collaboratively with state education agencies, released a set of criteria for states to use to evaluate and procure high-quality assessments.[1] The mathematics section of the document included five content-specific criteria to evaluate alignment of assessments to college- and career-ready mathematics standards:

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

In 2016, both the Thomas B. Fordham Institute and the Human Resources Research Organization (HumRRO) used the criteria to evaluate a set of statewide summative assessments: ACT Aspire, the Massachusetts Comprehensive Assessment System (MCAS), PARCC, and Smarter Balanced. Fordham[2] examined grades 5 and 8 assessments, while HumRRO[3] evaluated high school assessments. Reports for each of these studies included recommendations to improve the methodology. Achieve, in partnership with Student Achievement Partners and in consultation with other content and assessment experts, worked to make these improvements, and in 2018 used the updated methodology to review the ACT.[4]

This brief describes efforts to address the evaluation of two mathematics criteria—Assessing a balance of concepts, procedures, and applications (C.2) and Requiring a range of cognitive demand (C.4)—to provide conceptual and methodological clarity for states designing summative assessments, assessment developers, and organizations evaluating assessment alignment and quality. Both criteria C.2 and C.4 are interrelated. First, we describe Achieve's approach to addressing C.2: Assessing a balance of concepts, procedures, and applications, which Fordham in their review was not able to fully evaluate. Second, we examine C.4: Requiring a range of cognitive demand, which traditionally has used Webb's Depth of Knowledge (DOK) as its lens. Achieve proposes a new mathematics-specific approach to measuring cognitive complexity for C.4, and which draws on the methodology developed for C.2.

Introducing the Aspects of Rigor Matrix

Mathematics and assessment experts have long recognized the importance of assessing students' understanding of mathematical concepts, procedures, and applications. In 1993, the National Research Council and Mathematical Sciences Education Board noted:

Assessment is the means by which we determine what students know and can do. It tells teachers, students, parents, and policymakers something about what students have learned: the mathematical terms they recognize and can use,

[1] Council of Chief State School Officers. (2014). The Criteria for Procuring and Evaluating High-Quality Assessments. https://www.ccsso.org/resource-library/criteria-procuring-and-evaluating-high-quality-assessments

[2] Doorey, N., & Polikoff, M. (2016). Evaluating the content and quality of next generation assessments. Thomas B. Fordham Institute. https://edexcellence.net/publications/evaluating-the-content-and-quality-of-next-generation-assessments

[3] Schultz, S., Michaels, H., Dvorak, R. & Wiley, C. (2016). Evaluating the Content and Quality of Next Generation High School Assessments: Final Report. Human Resources Research Organization. https://www.humrro.org/corpsite/sites/default/files/HQAP_HumRRO_High_School_Study_Final%20Report.pdf

[4] Achieve. (2018). Independent Analysis of the Alignment of the ACT to the Common Core State Standards. https://www.achieve.org/achieve-act-review

the procedures they can carry out, the kind of mathematical thinking they do, the concepts they understand, and the problems they can formulate and solve. (p. 23)[5]

CCSSO incorporated this idea into criterion C.2: Assessing a balance of concepts, procedures, and applications. In operationalizing this criterion for evaluating assessment alignment and quality, Fordham drew on the methodology developed by the Center for Assessment[6] and instructed reviewers to categorize items by the predominant focus of an individual math item) — whether the item focused on a concept, procedure, or application. As Fordham described in its recommendations, this led to significant challenges resulting in not being able to provide a rating for criterion C.2:

> *First, the requirement of categorizing items by their predominant focus led to a failure to recognize and give credit for items that address two or more categories. Second, and related, items that were coded as measuring "combined" skills were not counted in any way, so assessments with more "combined" items were penalized in accordance with the tentative scoring guidance. Third, the broad definition of "application" (i.e., any item that includes a context) resulted in many items that also assessed conceptual understanding and/or procedural skill/fluency to be categorized as only application because they included use of a context (even if trivial). This resulted in a lowered rating and a failure to recognize the other competencies being addressed. (p. 77)*[7]

One of the key differences between Achieve's methodology and the original one used by Fordham and HumRRO was the development and implementation of the Aspects of Rigor Matrix (AOR) to evaluate 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. Achieve developed and used a new methodology to evaluate the balance of aspects of rigor, which allows items to be aligned to any combination of the aspects of rigor.

The matrix, shown below, permits an item to align to one or more aspects of rigor. Each item receives one of six designations; an item that addresses procedural skill and application, for example, would be recognized for both.

Table 1: Aspects of Rigor Matrix (AOR)

	The item does not involve application.[8]	The item involves an application.
The item targets procedural skill expected by the grade level.	**P**	**P-A**
The item targets conceptual understanding[9] 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.	**P-C**	**P-C-A**
The item targets conceptual understanding. Students may explain, strategize, evaluate, determine, compare, or classify.	**C**	**C-A**

Achieve has found several advantages in adopting the AOR Matrix for evaluating items' procedural skill, conceptual understanding, and application. First, this approach recognizes that individual mathematics items may incorporate multiple aspects of rigor. Second, this approach allows for a more fine-grained analysis of assessment items through the use of clear descriptors for each of the categories. Finally, we hope that the AOR Matrix provides assessment developers with more helpful and specific targets to achieve a balance of concepts, procedures, and applications.

[5] National Research Council. (1993). Measuring What Counts: A Conceptual Guide for Mathematics Assessment. Washington, DC: The National Academies Press. https://doi.org/10.17226/2235

[6] 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. http://www.nciea.org/sites/default/files/publications/Guide-to-Evaluating-CCSSO-Criteria-Test-Content_020316.pdf

[7] Doorey, N., & Polikoff, M. (2016). Evaluating the content and quality of next generation assessments. Thomas B. Fordham Institute. https://edexcellence.net/publications/evaluating-the-content-and-quality-of-next-generation-assessments

[8] Names or mathematical referents (e.g., units of measure) may be present in item, but should not be considered "application" in the sense intended by the shifts in CCR standards.

[9] Conceptual understanding refers to the mathematics used to respond to an item, not the complexity of the question itself.

A New Approach to Evaluate Cognitive Complexity in Mathematics

According to CCSSO criterion C.4, assessments should "require all students to demonstrate a range of higher-order, analytical thinking skills in mathematics based on the depth and complexity of college- and career-ready standards, allowing robust information to be gathered for students with varied levels of achievement."[10] The Fordham and HumRRO studies examined this through the lens of Webb's Depth of Knowledge (DOK), assigning DOK level ratings to items. In the DOK framework, an item that focuses on the recall of facts or definitions or on performing a simple algorithm would be considered DOK Level 1 (recall) while an item that goes beyond a habitual response and requires students to make some decisions would be at DOK Level 2 (skill/concept). Items that require planning and the use of evidence, beyond that expected in Levels 1 and 2, would be at DOK Level 3 (strategic thinking). Lastly, an item that requires complex reasoning over a period of time would be at DOK Level 4 (extended thinking). [11]

Given the importance of the three aspects of rigor and the emphasis of these aspects in current state standards,[12] it is reasonable to expect that the overall collection of items in a summative assessment require a range of demand within each of the areas of rigor. The DOK framework, however, does not have the ability to distinguish whether each aspect of rigor is represented at each of the DOK levels. To address this, Achieve has worked to establish a direct relationship between C.2 and C.4, where C.2 provides information on the balance of procedures, concepts, and applications, while C.4 addresses the levels of complexity within those aspects of rigor.

To this end, Achieve proposes a method for classifying the cognitive complexity of items according to the corresponding aspects of rigor. This method is based on the AOR Matrix shown in Table 1 and the Levels of Complexity shown in Table 2. A reviewer will begin an item analysis by considering which of the three aspects of rigor are targeted by an assessment item (i.e., C.2). Each item will align to either one, two, or three aspects of rigor and will occupy some cell in the AOR Matrix. Based on that analysis, a reviewer will then consider the complexity levels for each targeted aspect (i.e., C.4). An item that addresses all three aspects of rigor will also be assigned complexity levels for all three aspects. Similarly, an item that is tagged **P-C** in the AOR Matrix will be assigned procedural and conceptual complexity levels. If that **P-C** item aligns to the descriptors for Level 2 Procedural Complexity and Level 1 Conceptual Complexity the analysis might be recorded as:

Procedural	Conceptual	Application
2	1	—

Such a framework will permit an overall analysis of cognitive range in terms of the aspects of rigor, and to determine if each aspect of rigor is assessed at varied levels. Overall, a summative assessment should include a balance of rigor and a variation in the corresponding levels of complexity. Ideally, a summative assessment should include conceptual items and application items at complexity levels 2 and 3.

[10] Council of Chief State School Officers. (2014). The Criteria for Procuring and Evaluating High-Quality Assessments. https://www.ccsso.org/resource-library/criteria-procuring-and-evaluating-high-quality-assessments. p.13.

[11] Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education, 20(1), 7-25.

[12] Achieve (2017) Strong standards: A review of changes to state standards since the Common Core. https://www.achieve.org/files/StrongStandards.pdf

Table 2: Levels of Complexity

	Level 1	Level 2	Level 3
Procedural Complexity:[13]	Solving the problem entails little procedural[14] demand or procedural demand is below grade level.	Solving the problem entails common or grade-level procedure(s) with friendly numbers.	Solving the problem requires common or grade-level procedure(s) with unfriendly numbers,[15] an unconventional combination of procedures, or requires unusual perseverance or organizational skills in the execution of the procedure(s).
Conceptual Complexity:[16]	Solving the problem requires students to recall or recognize a grade-level concept. The student does not need to relate concepts or demonstrate a line of reasoning.	Students may need to relate multiple grade-level concepts or different types, create multiple representations or solutions, or connect concepts with procedures or strategies. The student must do some reasoning, but may not need to demonstrate a line of reasoning.	Solving the problem requires students to relate multiple grade-level concepts and to evidence reasoning, planning, analysis, judgment, and/or creative thought OR work with a sophisticated (nontypical) line of reasoning.
Application Complexity:	Solving the problem entails an application of mathematics, but the required mathematics is either directly indicated or obvious.	Solving the problem entails an application of mathematics and requires an interpretation of the context to determine the procedure or concept (may include extraneous information). The mathematics is not immediately obvious. Solving the problem requires students to decide what to do.	In addition to an interpretation of the context, solving the problem requires recognizing important features, and formulating, computing, and interpreting results as part of a modeling process.

Examples

The publicly released item examples below, selected from various sources, illustrate how the AOR Matrix and the Levels of Complexity work together to support the analysis of assessment items.

Example 1: This North Carolina Grade 6 item involves procedural skill in an application, so it is classified **P-A** in the AOR Matrix. In this item (administered without a calculator) the required computation is below grade level, so the Procedural Complexity is at Level 1. For application, the required mathematics is rather obvious at this grade level so the Application Complexity is also at level 1.[17]

Heather earns $8.00 per hour for walking a dog. How many hours must she work to earn $256.00?

Proc.	Conc.	Appl.
1	—	1

[13] This is based on the *NAEP States Item to Item Comparison Study*, NAEP Validity Studies Panel; Philip Daro, Gerunda Hughes, Sami Kamito, Fran Stancavage, Natalie Tucker-Bradway; American Institutes for Research, 2018.

[14] A procedure is a step by step sequence that can be memorized and executed without understanding or attending to the meaning of the quantities; a procedure is useful for a class of problems or situations. Computations that are likely to be known from memory are considered procedural.

[15] Unfriendly numbers: The student is likely to get the problem wrong not because of the targeted procedure but because of the numbers involved.

[16] This is based on the conceptual aspects of the *Mathematics Framework for the 2015 National Assessment of Educational Progress*; National Assessment Governing Board, 2014.

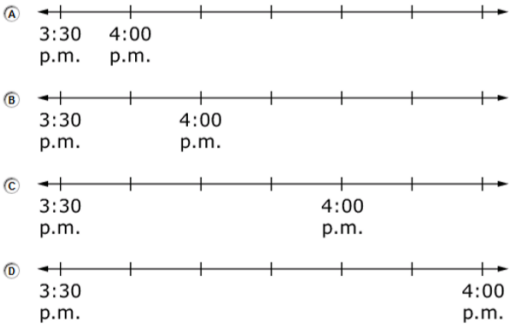
[17] This item is from the North Carolina READY End-of-Grade Assessment Mathematics Grade 6 Student Booklet, p. 7.

Example 2: This Smarter Balanced Grade 3 item targets conceptual understanding in an application, so it is classified **C-A** in the AOR Matrix. The Conceptual Complexity is Level 2 as students must connect concepts (connecting number line intervals to time) with a strategy. The mathematics in the application (Paul read for 45 minutes, starting at 3:30) is directly indicated, so the Application Complexity is Level 1.[18]

Paul made a number line to show the times he started reading and finished reading.

Paul read for 45 minutes.

Which number line shows 4:00 p.m. in the correct place on Paul's number line?



Proc.	Conc.	Appl.
—	2	1

Example 3: This Grade 8 Smarter Balanced item targets procedures, concepts and application, so it is classified **P-C-A** in the AOR Matrix. This is Level 3 Conceptual Complexity as students must demonstrate a sophisticated line of reasoning. This is Level 3 Procedural Complexity as the students work with an unconventional combination of procedures. Lastly, this is Level 2 Application Complexity as students must use the quantities in the context to determine the procedures and concepts to use.[19]

The surface area of Earth is 5.1×10^8 square km, and 71% of Earth's surface is covered in water. The diameter of the moon is 3475 km, and there is no water on the surface of the moon.

Which has more dry land, Earth or the moon? Show and/or explain the work necessary to support your answer.

Proc.	Conc.	Appl.
3	3	2

[18] This item is from the Smarter Balanced Assessment Consortium: Mathematics Practice Test Scoring Guide Grade 3, 08/01/2016, p. 18.

[19] This item is from the Smarter Balanced Scoring Guide For Selected Short-Text Mathematics Items (Field Test 2014), p. 17.

Example 4: This Grade 5 item from Illustrative Mathematics targets conceptual understanding, so it is classified **C** in the AOR Matrix. Students must use a sophisticated line of reasoning to compare the values, so the Conceptual Complexity is at Level 3.[20]

Decide which number is greater without multiplying.

a. 817 or 235×817

b. 99 or $\frac{1}{4} \times 99$

c. $\frac{51}{100}$ or $\frac{51}{100} \times 301$

d. $\frac{13}{90}$ or $\frac{2}{3} \times \frac{13}{90}$

e. $\frac{101}{102}$ or $\frac{101}{102} \times \frac{101}{102}$

f. $\frac{99}{5}$ or $\frac{99}{5} \times \frac{1}{2}$

g. $\frac{8}{21} \times 40$ or $\frac{28}{21} \times 40$

h. $\frac{8}{3} \times \frac{5}{7}$ or $\frac{8}{3} \times \frac{9}{4}$

Proc.	Conc.	Appl.
—	3	—

Example 5: This Smarter Balanced high-school level task is procedural and is classified **P** in the AOR Matrix. Since this is at grade level, as high school students are expected to rewrite expressions involving radicals and rational exponents, the Procedural Complexity is at Level 2.[21]

Select an expression that is equivalent to $\sqrt{3^8}$.

Ⓐ $3^{\frac{1}{4}}$

Ⓑ 3^3

Ⓒ 3^4

Ⓓ 3^6

Proc.	Conc.	Appl.
2	—	—

[20] This item is from the 5.NF.B.5 tasks at Illustrative Mathematics.

[21] This item is from the Smarter Balanced Assessment Consortium: Mathematics Practice Test Scoring Guide High School, 08/01/2016, p. 4.

Summary

The tools in this document provide a new way to think about and analyze mathematical rigor in summative assessments. The analysis is done in two steps: (1) determining the aspects of rigor targeted in an assessment item, and (2) determining the levels of complexity for each of the targeted aspects. The tools provide a lens by which the balance of rigor and cognitive demand in an assessment may be evaluated. Ideally, a summative assessment will reflect a balance in the aspects of rigor and cover a range of levels of complexity. At a minimum, assessments should include items that reach Levels 2 and 3 for both Conceptual and Application Complexity. However, we recommend caution with items at Procedural Complexity Level 3, as such items may require excessive time and may not provide specific and useful information.

We believe this new approach to evaluating item-level cognitive complexity has important advantages over traditional approaches like DOK. This model provides specific and helpful feedback on item complexity and it does so through the common language of procedures, concepts, and applications. Additionally, this approach creates a more seamless system for assessment review by directly connecting the work in CCSSO criterion C.2 to the expectations in criterion C.4.

Achieve is grateful to the Bill & Melinda Gates Foundation, the William and Flora Hewlett Foundation, and the Charles and Lynn Schusterman Family Foundation for their generous support of this work.

APPENDIX A: THE TOOLS

The Aspects of Rigor Matrix

	The item does not involve application.[22]	The item involves an application.
The item targets procedural skill expected by the grade level.	**P**	**P-A**
The item targets conceptual understanding[23] 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.	**P-C**	**P-C-A**
The item targets conceptual understanding. Students may explain, strategize, evaluate, determine, compare, or classify.	**C**	**C-A**

The Levels of Complexity

	Level 1	Level 2	Level 3
Procedural Complexity:[24]	Solving the problem entails little procedural[25] demand or procedural demand is below grade level.	Solving the problem entails common or grade-level procedure(s) with friendly numbers.	Solving the problem requires common or grade-level procedure(s) with unfriendly numbers,[26] an unconventional combination of procedures, or requires unusual perseverance or organizational skills in the execution of the procedure(s).
Conceptual Complexity:[27]	Solving the problem requires students to recall or recognize a grade-level concept. The student does not need to relate concepts or demonstrate a line of reasoning.	Students may need to relate multiple grade-level concepts or different types, create multiple representations or solutions, or connect concepts with procedures or strategies. The student must do some reasoning, but may not need to demonstrate a line of reasoning.	Solving the problem requires students to relate multiple grade-level concepts and to evidence reasoning, planning, analysis, judgment, and/or creative thought OR work with a sophisticated (nontypical) line of reasoning.
Application Complexity:	Solving the problem entails an application of mathematics, but the required mathematics is either directly indicated or obvious.	Solving the problem entails an application of mathematics and requires an interpretation of the context to determine the procedure or concept (may include extraneous information). The mathematics is not immediately obvious. Solving the problem requires students to decide what to do.	In addition to an interpretation of the context, solving the problem requires recognizing important features, and formulating, computing, and interpreting results as part of a modeling process.

Note: A standard may address fluency with understanding, but corresponding assessment items that do not require some evidence of understanding should be coded as procedural, as students may proceduralize the understandings at any time.

[22] Names or mathematical referents (e.g., units of measure) may be present in item, but should not be considered "application" in the sense intended by the shifts in CCR standards.
[23] Conceptual understanding refers to the mathematics used to respond to an item, not the complexity of the question itself.
[24] This is based on the *NAEP States Item to Item Comparison Study*, NAEP Validity Studies Panel; Philip Daro, Gerunda Hughes, Sami Kamito, Fran Stancavage, Natalie Tucker-Bradway; American Institutes for Research, 2018.
[25] A procedure is a step by step sequence that can be memorized and executed without understanding or attending to the meaning of the quantities; a procedure is useful for a class of problems or situations. Computations that are likely to be known from memory are considered procedural.
[26] Unfriendly numbers: The student is likely to get the problem wrong not because of the targeted procedure but because of the numbers involved.
[27] This is based on the conceptual aspects of the *Mathematics Framework for the 2015 National Assessment of Educational Progress*; National Assessment Governing Board, 2014.