Evaluating the Content and Quality of Next Generation Assessments

Legislative Education Staff Network (LESN) Seminar
Amber Northern
SVP for Research

June 29, 2016

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

The Fordham Team

Amber Northern

Senior VP for Research,
Thomas B. Fordham Institute

Victoria Sears

Research Manager,
Thomas B. Fordham Institute

Charles Perfetti

ELA/Literacy Content Lead and
Distinguished Professor of Psychology
at the University of Pittsburgh

Nancy Doorey

Educational consultant with
assessment-policy expertise

Morgan Polikoff

Assistant Professor at the
University of Southern California
and expert in alignment methods

Roger Howe

Math Content Lead and Professor
of Mathematics at Yale University

Study Components

Phase 1

- Item Review: Test Forms
- Generalizability (Document) Review: Blueprints, assessment frameworks, etc. (subset of item reviewers)
- Accessibility Review: Separate panel (joint review with HumRRO)

Phase 2

- Aggregation of Item Review and Generalizability Results and development of consensus statements

Key Study Questions

1. Do the assessments place strong emphasis on the most important content for college and career readiness (CCR) as called for by the Common Core State Standards and other CCR standards? **(Content)**

2. Do they require all students to demonstrate the range of thinking skills, including higher-order skills, called for by those standards? **(Depth)**

3. What are the overall strengths and weaknesses of each assessment relative to the examined criteria for ELA/Literacy and mathematics? **(Overall Strengths and Weaknesses)**

Review Panels and Design

- We received **over 200 reviewer recommendations** from various assessment and content experts and organizations, as well as each of the four participating assessment programs.
- In vetting applicants, we prioritized extensive content and/or assessment expertise, deep familiarity with the CCSS, and prior experience with alignment studies. Not eligible: employees of test programs or writers of the standards
- Final review panels were comprised of classroom educators, content experts, and experts in assessment and accessibility. We included at least one reviewer recommended by each participating program on each panel.
- **Seven test forms were reviewed per grade level and content area** (2 forms each for Smarter Balanced, PARCC, and ACT Aspire, and 1 form for MCAS). Reviewers were randomly assigned to forms using a ñjigsawò approach across testing programs to minimize major differences across panels and enhance inter-rater reliability.

Council of Chief State School Officers (CCSSO) Criteria Evaluated

A. Meet Overall Assessment Goals and Ensure Technical Quality

A.5 Providing accessibility to all students, including English learners and students with disabilities (*HumRRO report only*)

B. Align to Standards – English Language Arts/Literacy

B.1 Assessing student reading and writing achievement in both ELA and literacy

B.2 Focusing on complexity of texts

B.3 Requiring students to read closely and use evidence from texts

B.4 Requiring a range of cognitive demand

B.5 Assessing writing

B.6 Emphasizing vocabulary and language skills

B.7 Assessing research and inquiry

B.8 Assessing speaking and listening *(measured but not counted)*

B.9 Ensuring high-quality items and a variety of item types

C. Align to Standards – Mathematics

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

Content criteria: Orange

Depth criteria: Blue

Ratings

Excellent Match
Good Match
Limited/Uneven Match
Weak Match

Ratings Tally by Program

ELA/Literacy Ratings Tally by Program

ACT Aspire	E	G	G	G	L	L	L	W	W
MCAS	E	G	G	G	L	L	W	W	W
PARCC	E	E	E	E	E	E	G	G	W
Smarter Balanced	E	E	E	E	G	G	G	G	L

Mathematics Ratings Tally by Program

ACT Aspire	E	E	L	W
MCAS	E	E	E	L
PARCC	E	G	G	G
Smarter Balanced	E	G	G	L

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

Overall Content and Depth Ratings for ELA/Literacy and Mathematics

	ACT Aspire	MCAS	PARCC	Smarter Balanced
ELA/Literacy CONTENT	L	L	E	E
ELA/Literacy DEPTH	G	G	E	G
Mathematics CONTENT	L	L	G	G
Mathematics DEPTH	G	E	G	G

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

ELA/Literacy Content Ratings Summary

Criteria	ACT Aspire	MCAS	PARCC	Smarter Balanced
I. CONTENT: Assesses the content most needed for College and Career Readiness	L	L	E	E
B.3 Reading:* Tests require students to read closely and use specific evidence from texts to obtain and defend correct responses.	L	G	E	E
B.5 Writing:* Tasks require students to engage in close reading and analysis of texts. Across each grade band, tests include a balance of expository, persuasive/argument, and narrative writing.	L	W	E	E
B.6 Vocabulary and language skills: Tests place sufficient emphasis on academic vocabulary and language conventions as used in real-world activities.	G	L	E	G
B.7 Research and inquiry: Assessments require students to demonstrate the ability to find, process, synthesize, and organize information from multiple sources.	L	W	E	E
B.8 Speaking and listening: Over time, and as assessment advances allow, the assessments measure speaking and listening communication skills.**	W	W	W	L

* The criteria recommended to be more heavily emphasized are underlined.

** The methodology indicates that criterion B.8 (speaking and listening) should be included "over time, and as assessment advances allow." Thus B.8 ratings are not included in the overall rating for Content.

*** The criterion B.2 rating is based solely on program documentation, as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the criterion B.2 rating as heavily when deciding the overall depth rating.

LEGEND E Excellent Match · G Good Match · L Limited/Uneven Match · W Weak Match

Cells for which the ratings are not used in determining Content and Depth ratings

ELA/Literacy Depth Ratings Summary

Criteria	ACT Aspire	MCAS	PARCC	Smarter Balanced
II. DEPTH: Assesses the depth that reflects the demands of College and Career Readiness	G	G	E	G
B.1 Text quality and types:* Tests include an aligned balance of high-quality literary and informational texts.	G	G	G	E
B.2 Complexity of texts:* Test passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.***	G	G	G	G
B.4 Cognitive demand: The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.	W	L	E	G
B.9 High-quality items and variety of item types: Items are of high technical and editorial quality and test forms include at least two item types with at least one that requires students to generate a response.	E	E	E	G

* The criteria recommended to be more heavily emphasized are underlined.

** The methodology indicates that criterion B.8 (speaking and listening) should be included "over time, and as assessment advances allow." Thus B.8 ratings are not included in the overall rating for Content.

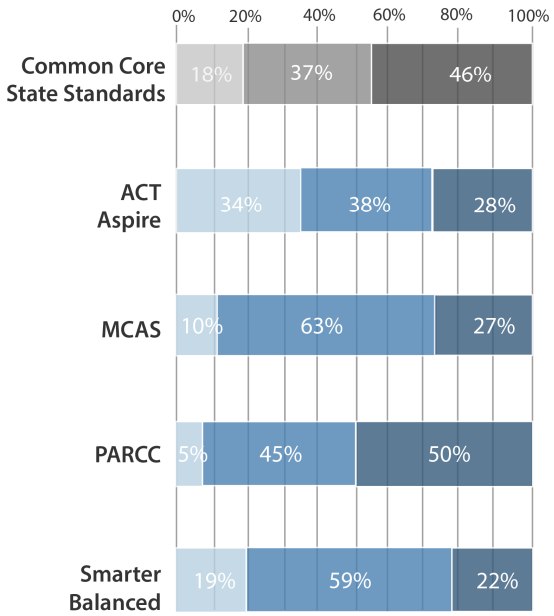
*** The criterion B.2 rating is based solely on program documentation, as reviewers were not able to rate the extent to which quantitative measures are used to place each text in a grade band. Thus, reviewers did not consider the criterion B.2 rating as heavily when deciding the overall depth rating.

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

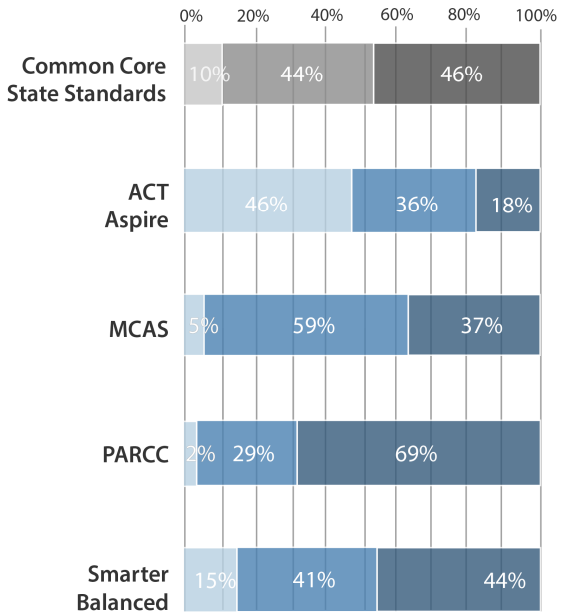
Cells for which the ratings are not used in determining Content and Depth ratings

Criterion B.4 Findings: The Distribution of Cognitive Demand in ELA/Literacy

ELA/Literacy Grade 5



ELA/Literacy Grade 8



Legend

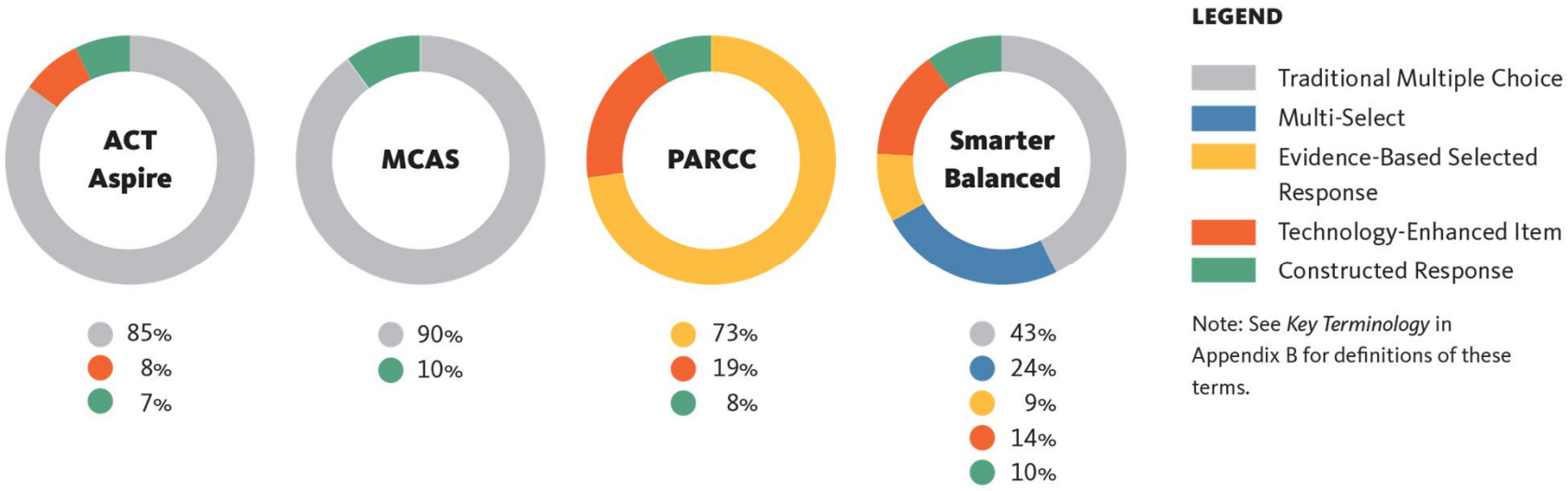
Level 1 includes basic recall of facts, concepts, information, or procedures.

Level 2 includes skills and concepts, such as the use of information (graphs) or requires two or more steps with decision points along the way.

Levels 3 and 4 include short-term strategic thinking, extended thinking, and often the application of concepts. Levels 3 and 4 are also referred to as "higher-order thinking skills."

Note: Percentages in the table represent percentages of score points at each DOK level. Results for a particular grade and program were generated by averaging across all raters and forms for that grade and program.

Criterion B.9 Findings: Distribution of Item Types in ELA/Literacy Tests



Mathematics Content Ratings Summary

Criteria	ACT Aspire	MCAS	PARCC	Smarter Balanced
I. CONTENT: Assesses the content most needed for College and Career Readiness	L	L	G	G
C.1 Focus:* Tests focus strongly on the content most needed in each grade or course for success in later mathematics (i.e., major work).	W	L	G	G
C.2: Concepts, procedures, and applications: Assessments place balanced emphasis on the measurement of conceptual understanding, fluency and procedural skill, and the application of mathematics.**	–	–	–	–

* The criteria recommended to be more heavily emphasized are underlined.

** Both programs require, in their program documentation, the assessment of conceptual understanding, procedural skill/fluency, and application, although most do not clearly distinguish between procedural skill/fluency and conceptual understanding. Also, specific balance across these three types is not required. Due to variation across reviewers in how this criterion was understood and implemented, final ratings could not be determined with confidence. Therefore, for criterion C.2, only qualitative observations are provided for grades 5 and 8. (See Section I, *Findings* for more information.)

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

– Cells for which no quantitative rating could be determined

Mathematics Depth Ratings Summary

Criteria	ACT Aspire	MCAS	PARCC	Smarter Balanced
II. DEPTH: Assesses the depth that reflects the demands of College and Career Readiness	G	E	G	G
C.3 Connecting practice to content:* Test questions meaningfully connect mathematical practices and processes with mathematical content.	E	E	E	E
C.4 Cognitive demand: The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.	L	E	G	G
C.5 High-quality items and variety of item types: Items are of high technical and editorial quality and test forms include at least two item types, at least one that requires students to generate a response.	E	E	G	L

* The criteria recommended to be more heavily emphasized are underlined.

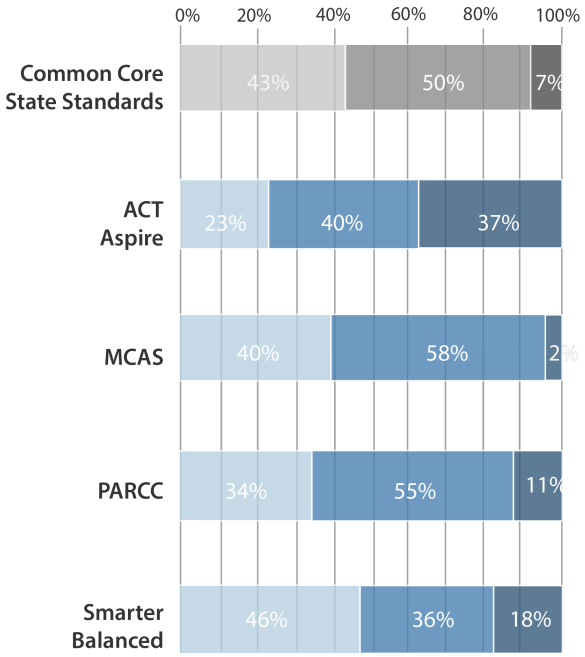
** Both programs require, in their program documentation, the assessment of conceptual understanding, procedural skill/fluency, and application, although most do not clearly distinguish between procedural skill/fluency and conceptual understanding. Also, specific balance across these three types is not required. Due to variation across reviewers in how this criterion was understood and implemented, final ratings could not be determined with confidence. Therefore, for criterion C.2, only qualitative observations are provided for grades 5 and 8. (See Section I, *Findings* for more information.)

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

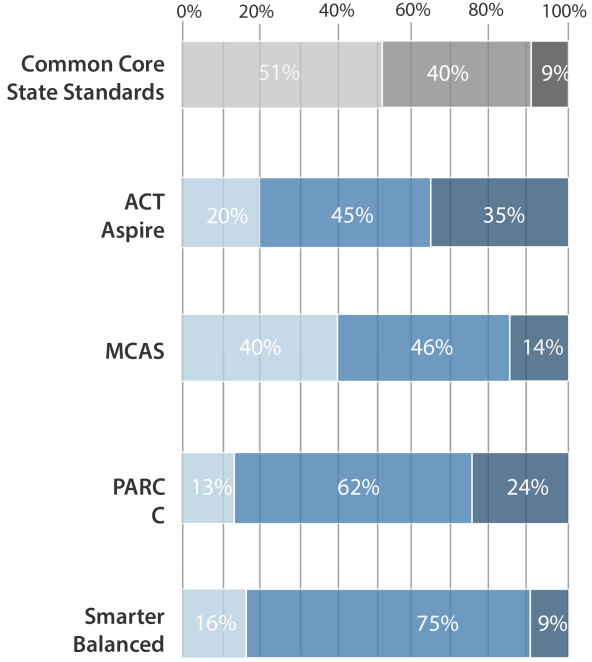
— Cells for which no quantitative rating could be determined

Criterion C.4 Findings: The Distribution of Cognitive Demand in Mathematics

Mathematics Grade 5



Mathematics Grade 8



Legend

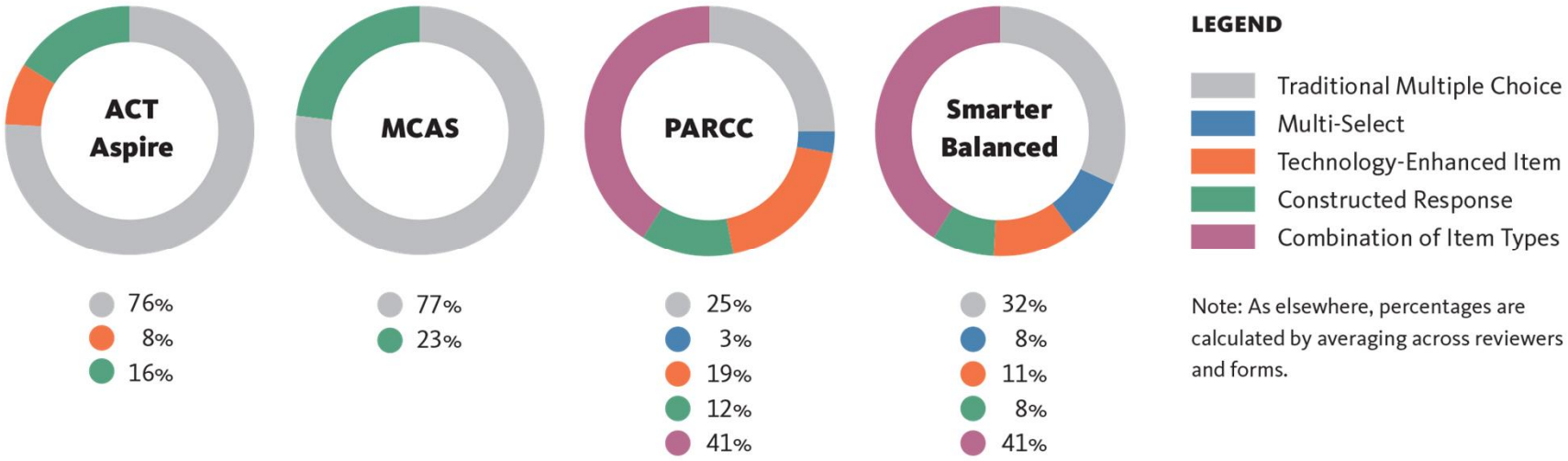
Level 1 includes basic recall of facts, concepts, information, or procedures.

Level 2 includes skills and concepts, such as the use of information (graphs) or requires two or more steps with decision points along the way.

Levels 3 and 4 include short-term strategic thinking, extended thinking, and often the application of concepts. Levels 3 and 4 are also referred to as "higher-order thinking skills."

Note: Percentages in the table represent percentages of score points at each DOK level. Results for a particular grade and program were generated by averaging across all raters and forms for that grade and program.

Criterion C.5 Findings: Distribution of Item Types in Mathematics Tests



Note: As elsewhere, percentages are calculated by averaging across reviewers and forms.

Slide 17

3 Awaiting the return of this graph from the designer with revisions.
Jonathan Lutton,

Sample: PARCC Program Strengths and Areas for Improvement (Grades 5/8)

Strengths ELA/Literacy

- Includes suitably complex texts
- Requires a range of cognitive demand
- Demonstrates variety in item types
- Requires close reading
- Assesses writing to sources, research, and inquiry
- Emphasizes vocabulary and language skills

Strengths Mathematics

- Reasonably well aligned to the priority content at each grade level
- Includes a distribution of cognitive demand that is similar to that of the standards at grade 5

Areas for Improvement ELA/Literacy

- Use of more research tasks requiring students to use multiple sources
- Developing the capacity to assess speaking and listening skills

Areas for Improvement Mathematics

- Further focus on the prioritized content at grade 5
- Addition of more items at grade 8 that assess standards at DOK 1
- Increased attention to accuracy of the itemsð primarily editorial, but in some instances mathematical

Closing Thoughts

Life--and tests--are full of tradeoffs:

- Testing time
- Cost
- Autonomy
- The unique factor
- Comparability

Thank you for your time

anorthern@edexcellence.net

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE