Evaluating the Content and Quality of Next Generation Assessments – Grades 5 & 8

CO Interim ESSA Committee Meeting
August 31, 2016

Victoria Sears
Research Manager, Thomas B. Fordham Institute

Slide 1

6 ok let's keep it for her but take it out of this presentation.
Amber Northern,

2 This slide can be removed easily enough, provided that the funders are familiar with our mission.

(I believe this content will prove useful for Alyssa's press release / webinar, however, so let's hold onto it for the time being.)
Jonathan Lutton,

5 again lets just do b7 and C4
Amber Northern,

2 I suggest changing the slide title to "Sample ELA Ratings and Consensus Statements (Criteria B.5)" - as well as for other examples on subsequent slides

Reading might make a better example for a couple of reasons (folks talk CCSS's reading expectations a lot, ACT had issues with our interpretation of B5 etc)

Also, the summary scores and statements are super small! Can you increase the font size for those, and make the grey box a lot smaller?
Victoria Sears,

4 or not...I keep seeing slides that I like...
Amber Northern,

3 I guess we could put B1 in too since we have more ELA criterion anyway, so we'll have 3 criterion total to call out specifically.
Amber Northern,

1 Same suggested edits as last slide :) Also think B1 would be a great example
Victoria Sears,

2 lets just do b7 and C4 since those are the ones we are highlighting earlier--so you can take this one out
Amber Northern,

1 ok keep this too. I'll look at this whole presentation with fresh eyes once you have made the changes and we can revisit things as needed...
Amber Northern,

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 and English 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

Project Partners

Center for Assessment

Developed and published the content alignment methodology

Thomas B. Fordham Institute
Advancing Educational Excellence

Implemented methodology (grades 5 & 8)

HumRRO
Human Resources Research Organization

Implemented methodology (high school)

Student Achievement Partners

Conducted reviewer training (TBFI)

HQAP | High-Quality Assessment Project

Funders supporting the study

Study Overview

- This study evaluates the **content, quality, and accessibility** of assessments for grades 5, 8, and high school for both mathematics and English language arts (ELA/Literacy)

- Evaluation criteria drawn from the content-specific portions of the Council of Chief State School Officers' (CCSSO's) "Criteria for Procuring and Evaluating High Quality Assessments"

- Aims to inform educators, parents, policymakers and other state and local officials of the strengths and weaknesses of several new next-generation assessments on the market (**ACT Aspire, PARCC, Smarter Balanced**)—as well as how a respected state test (**MCAS**) stacks up

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)**

4. Are the assessments accessible to all students, including English learners (ELs) and students with disabilities (SWDs)? (**Accessibility)**

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

Study Components

Phase 1

- Item Review: *Operational* Items and Test Forms
- Generalizability (Document) Review: Blueprints, assessment frameworks, etc. (subset of item reviewers)

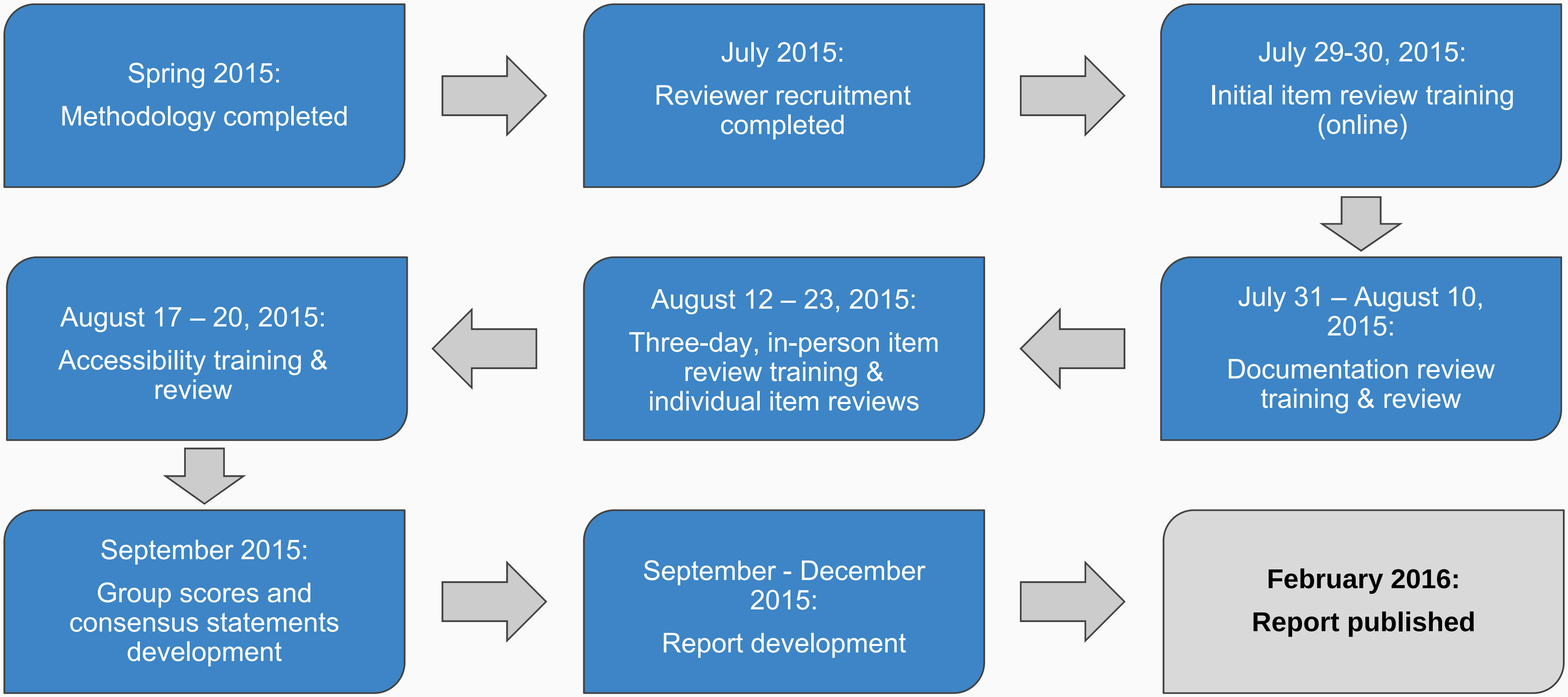
Phase 2

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

Phase 3

- Accessibility (joint review with HUMRRO)
  - Exemplar Review: Operational or Sample Items
  - Generalizability (Document) Review: Accessibility and Assessment frameworks, etc.

Study Timeline



Spring 2015:

Methodology completed

July 2015:

Reviewer recruitment completed

July 29-30, 2015:

Initial item review training (online)

July 31 – August 10, 2015:

Documentation review training & review

August 12 – 23, 2015:

Three-day, in-person item review training & individual item reviews

August 17 – 20, 2015:

Accessibility training & review

September 2015:

Group scores and consensus statements development

September - December 2015:

Report development

February 2016:

Report published

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 (n=8) were comprised of classroom educators, content experts, and experts in assessment and accessibility.
  - Fordham 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).
  - Fordham randomly assigned reviewers to forms using a "jigsaw" approach across testing programs
  - HumRRO randomly assigned reviewers to programs

Review activities, online vs. in-person

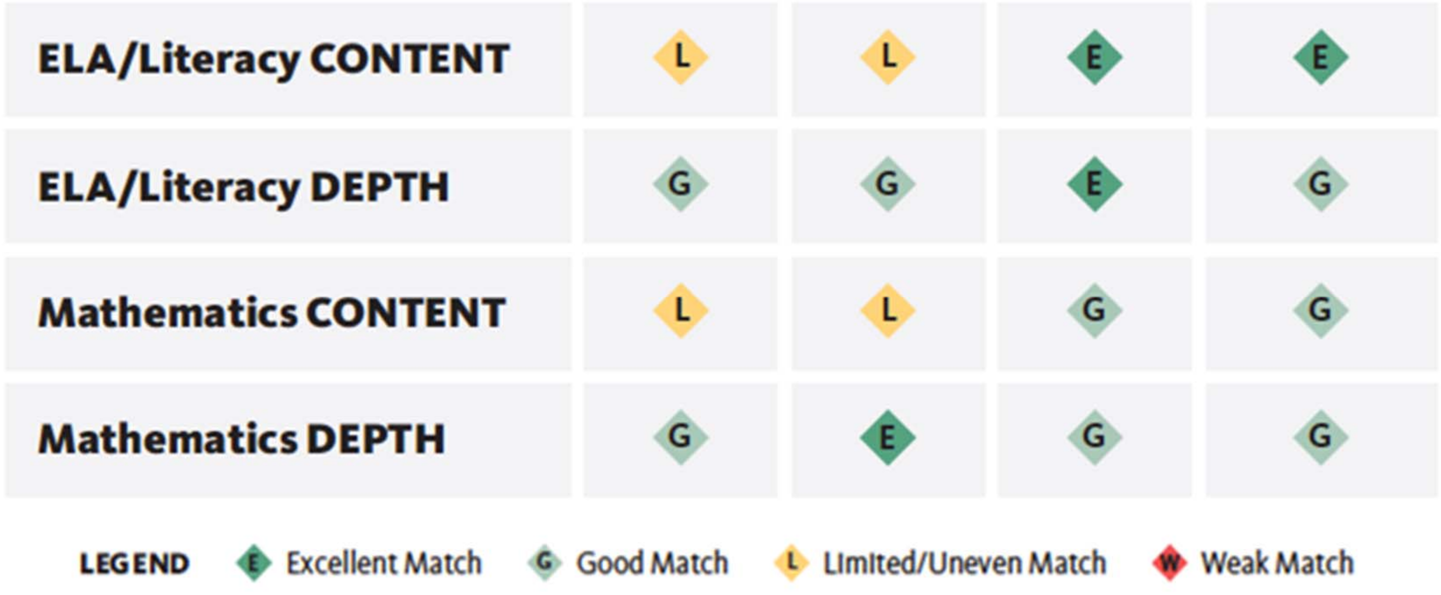
- Initial Reviewer Training(s) – study overview and introductions to CCSSO Criteria and individual testing platforms (online)
- In-person Training - connecting CCSSO criteria to study methodology, reviewer callibration, and commencing individual reviews (3 days, in-person)
- Phase 1 – Item review and Generalizability review (online)
- Phase 2 – Development of final scores & consensus statements (subset of item reviewers, online)
- Accessibility training & review (in-person, separate panel, joint with HumRRO)

Rating Labels

Each panel reviewed the ratings from the test forms, considered the results of the documentation review, and came to consensus on the criterion's rating--assigning the programs a rating on each of the ELA/Literacy and mathematics criterion:

- ○ **Excellent Match**
- ○ **Good Match**
- ○ **Limited/Uneven Match**
- ○ **Weak Match**

Overall Content and Depth Ratings for Grades 5 and 8 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

High-level findings (grades 5 & 8)

- Only PARCC & Smarter Balanced earned an EXCELLENT or GOOD MATCH to CCSSO criteria for high-quality assessments for both subjects
- While ACT Aspire and MCAS fared well regarding the overall quality of their test items and DEPTH assessed, these programs do not adequately assess some of the priority CONTENT in both subjects at one or both grades reviewed in the study

Final Ratings Tally by Subject and Program (Fordham Study, Grades 5 & 8)

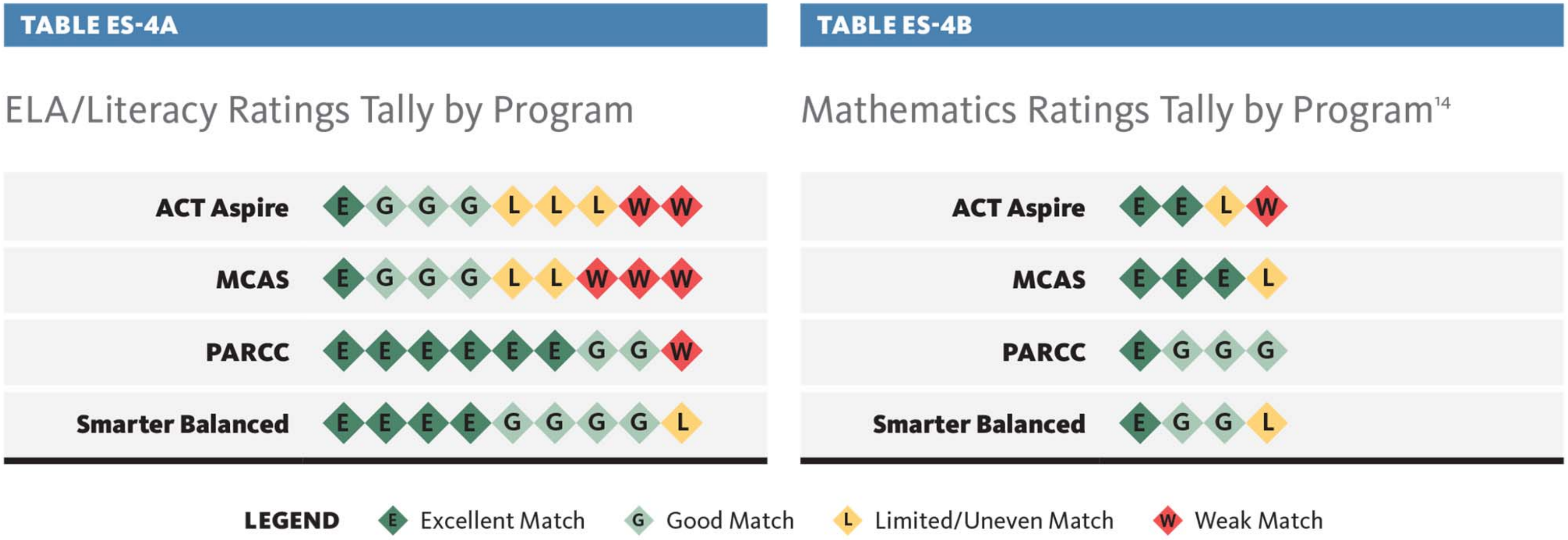


TABLE ES-4A

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

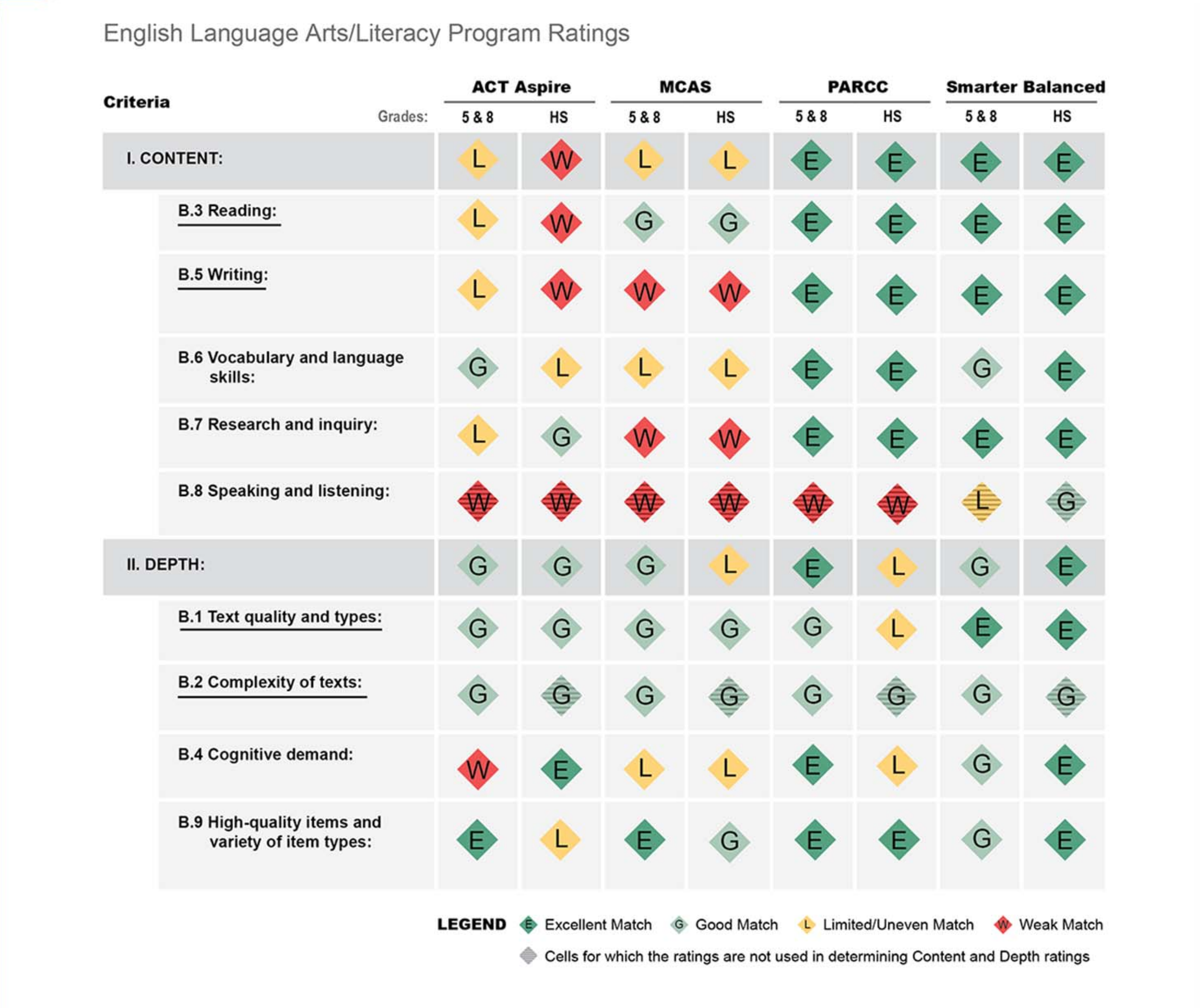
TABLE ES-4B

Mathematics Ratings Tally by Program[14]

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

Final Program Ratings – ELA (all grades)



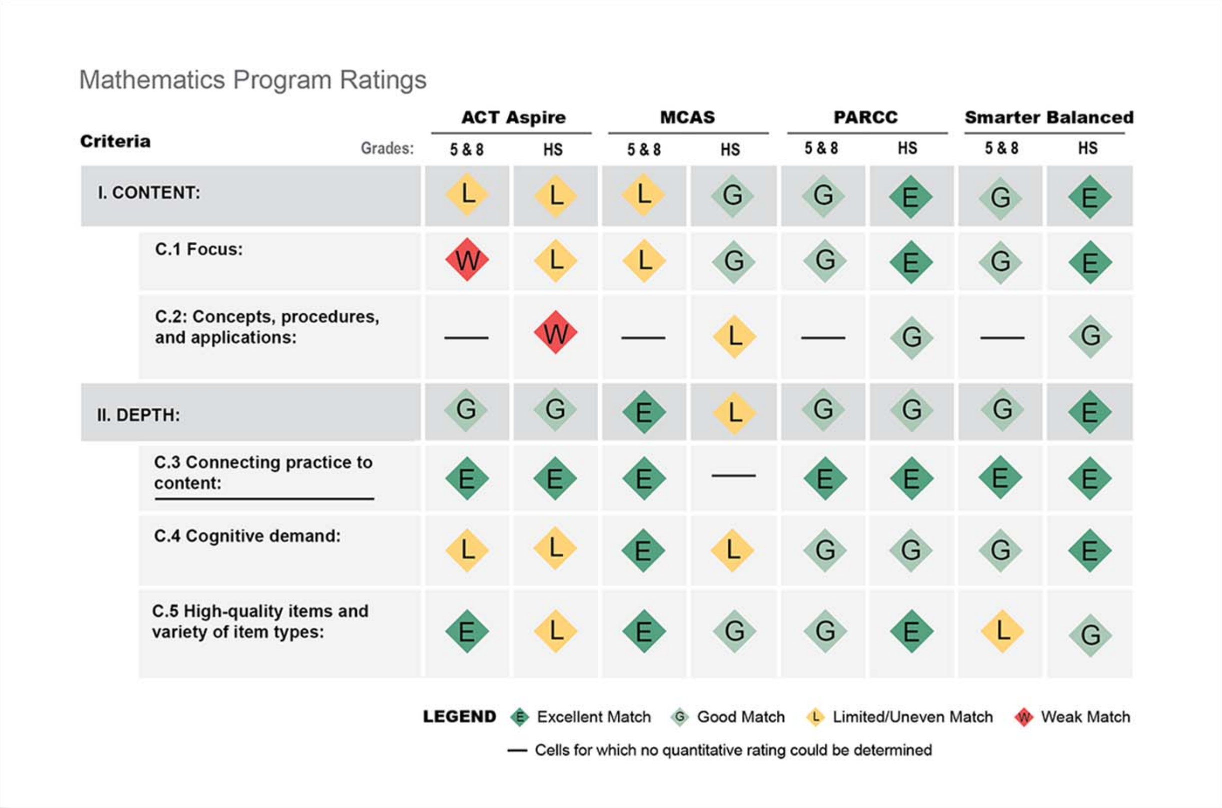
English Language Arts/Literacy Program Ratings

Criteria		ACT Aspire		MCAS		PARCC		Smarter Balanced	
	Grades:	5 & 8	HS	5 & 8	HS	5 & 8	HS	5 & 8	HS
I. CONTENT:		L	W	L	L	E	E	E	E
B.3 Reading:		L	W	G	G	E	E	E	E
B.5 Writing:		L	W	W	W	E	E	E	E
B.6 Vocabulary and language skills:		G	L	L	L	E	E	G	E
B.7 Research and inquiry:		L	G	W	W	E	E	E	E
B.8 Speaking and listening:		W	W	W	W	W	W	L	G
II. DEPTH:		G	G	G	L	E	L	G	E
B.1 Text quality and types:		G	G	G	G	G	L	E	E
B.2 Complexity of texts:		G	G	G	G	G	G	G	G
B.4 Cognitive demand:		W	E	L	L	E	L	G	E
B.9 High-quality items and variety of item types:		E	L	E	G	E	E	G	E

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

Final Program Ratings – Math (all grades)



Mathematics Program Ratings

Criteria		ACT Aspire		MCAS		PARCC		Smarter Balanced	
	Grades:	5 & 8	HS	5 & 8	HS	5 & 8	HS	5 & 8	HS
I. CONTENT:		L	L	L	G	G	E	G	E
C.1 Focus:		W	L	L	G	G	E	G	E
C.2: Concepts, procedures, and applications:		—	W	—	L	—	G	—	G
II. DEPTH:		G	G	E	L	G	G	G	E
C.3 Connecting practice to content:		E	E	E	—	E	E	E	E
C.4 Cognitive demand:		L	L	E	L	G	G	G	E
C.5 High-quality items and variety of item types:		E	L	E	G	G	E	L	G

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

— Cells for which no quantitative rating could be determined

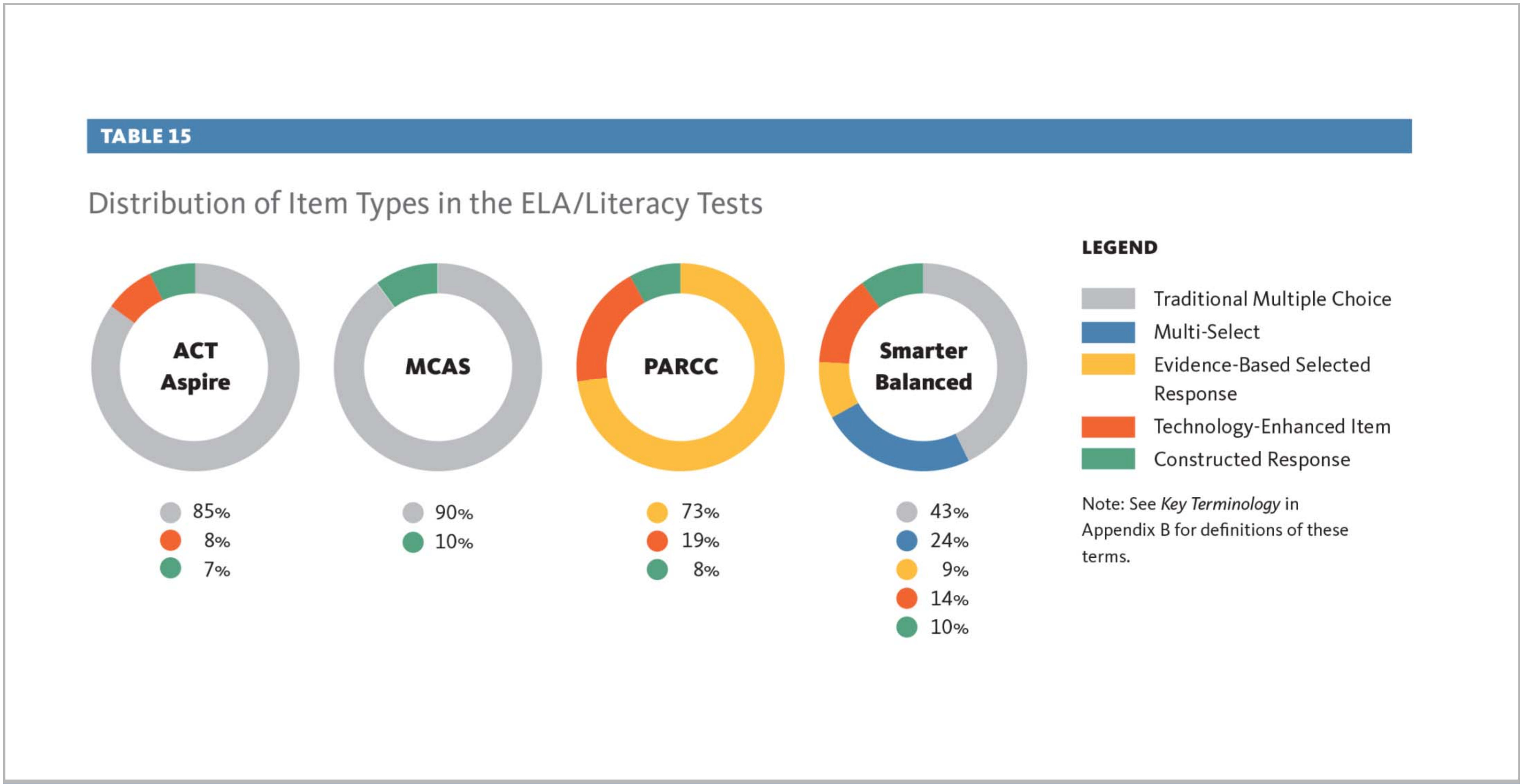


TABLE 15

Distribution of Item Types in the ELA/Literacy Tests

LEGEND

- ⬜ Traditional Multiple Choice
- 🟦 Multi-Select
- 🟨 Evidence-Based Selected Response
- 🟧 Technology-Enhanced Item
- 🟩 Constructed Response

Note: See *Key Terminology* in Appendix B for definitions of these terms.

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

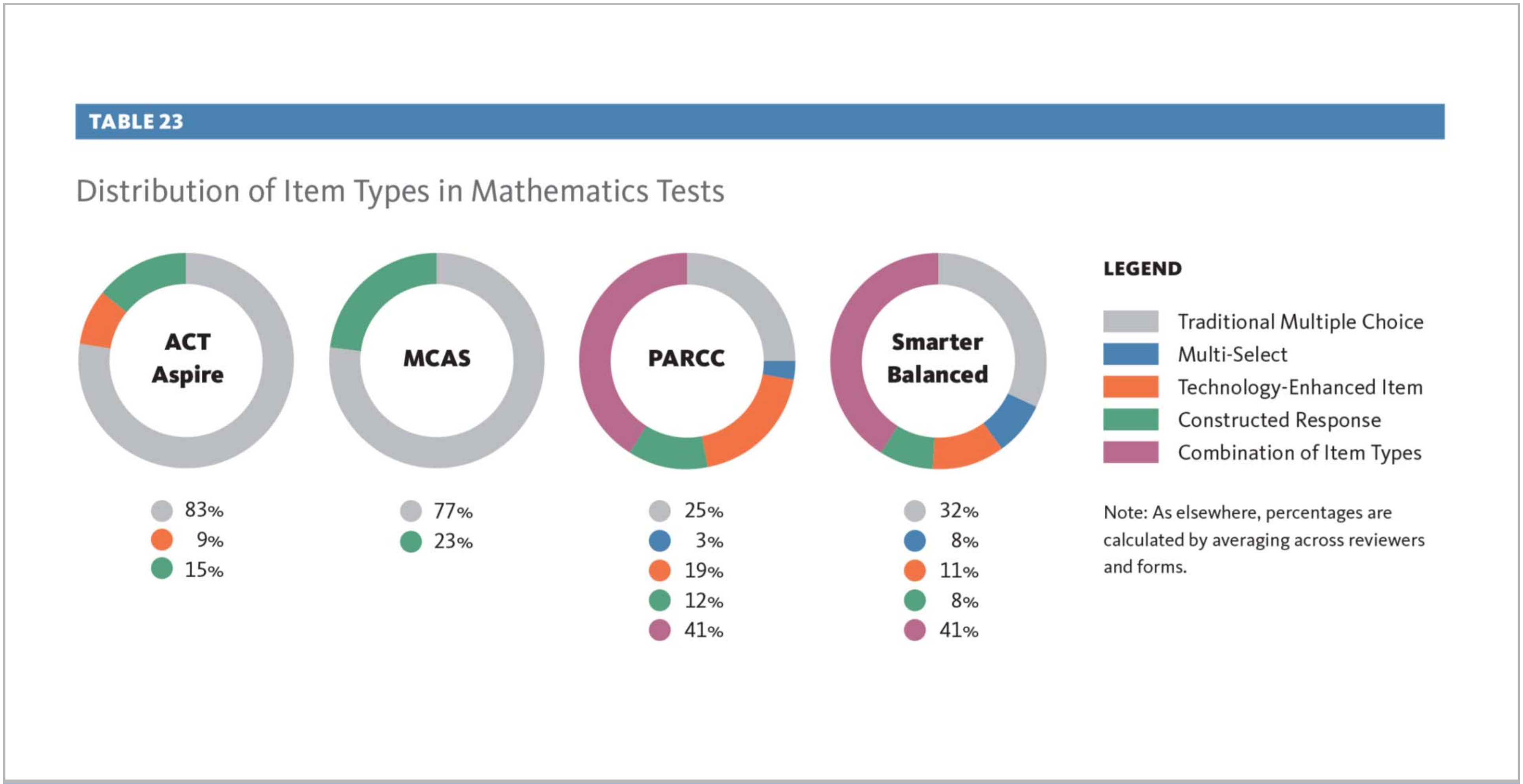


TABLE 23

Distribution of Item Types in Mathematics Tests

LEGEND

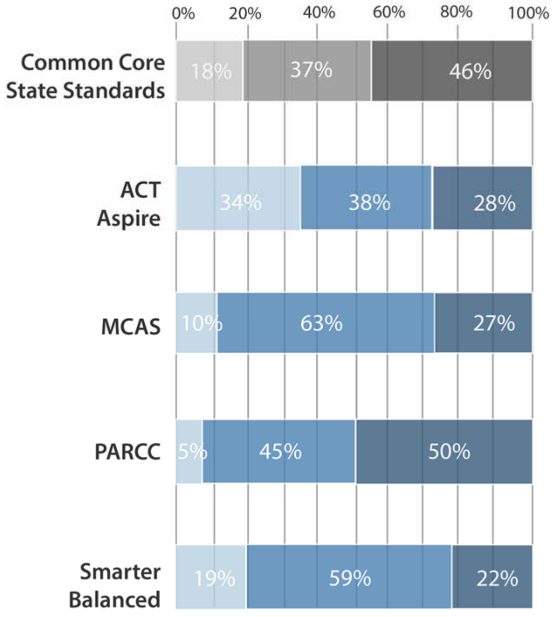
- ▨ Traditional Multiple Choice
- ▨ Multi-Select
- ▨ Technology-Enhanced Item
- ▨ Constructed Response
- ▨ Combination of Item Types

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

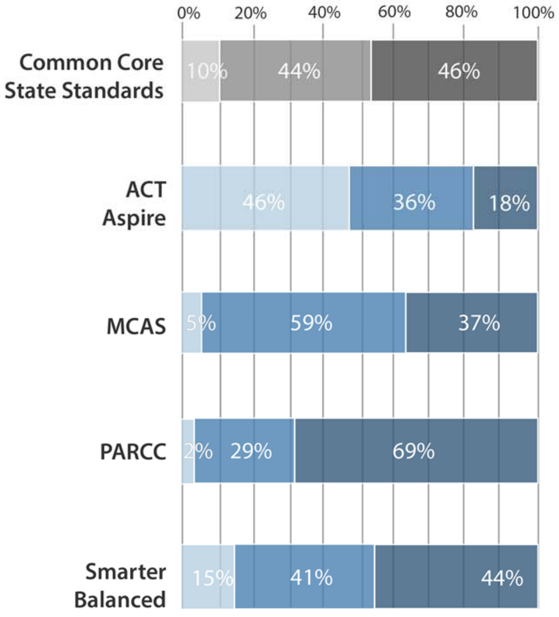
Criterion C.9 - Distribution of Item Types in Mathematics Tests

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

ELA/Literacy Grade 5



ELA/Literacy Grade 8



Legend

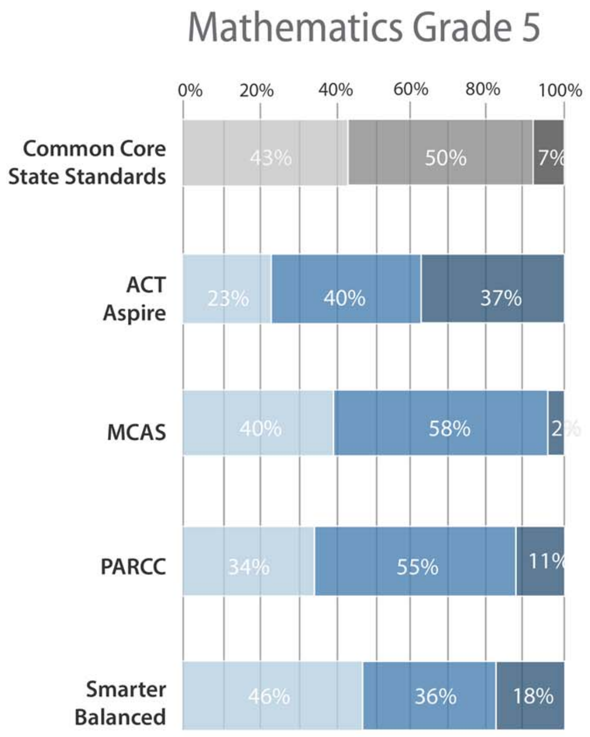
◇ **Level 1** includes basic recall of facts, concepts, information, or procedures.

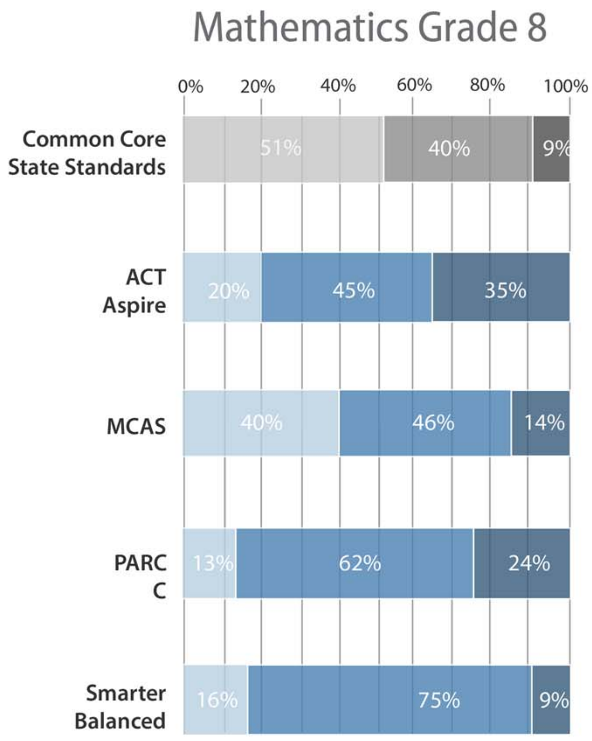
◆ **Level 2** includes skills and concepts, such as the use of information (graphs) or requires two or more steps with decision points along the way.

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





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.

Key Findings in ELA/Literacy

- Nearly all PARCC and Smarter Balanced reading items require close reading and analysis; smaller proportions for MCAS and ACT.
- PARCC and Smarter Balanced writing items require writing to sources. MCAS and ACT items do not (and writing not assessed at every grade on MCAS).
- PARCC items have the strongest match to the DOK of the standards. ACT items have the weakest match.
- All programs have at least good quality items. ACT, PARCC, and MCAS are excellent quality. ACT and MCAS are more reliant on traditional multiple choice items.

Key Findings in Mathematics

- PARCC and Smarter Balanced had a good match to the major work of the grade. MCAS had a more limited match, and ACT Aspire had a weak match.

- MCAS had the strongest match to the DOK of the standards. ACT's DOK greatly exceeded that of the standards, while PARCC and Smarter Balanced had more minor differences with the DOK of the standards.

- ACT and MCAS has excellent item quality. PARCC had some items with minor editorial and technical issues, but still received a good rating. Reviewers found more issues with Smarter Balanced item quality, including repeated items and quality issues on an average of 1-2 items per form.

Key Findings: Accessibility

- ACT Aspire has the fewest number of accessibility features (about 30 for their online assessment and about 35 for their paper-pencil assessment). PARCC and Smarter Balanced have over 50 features listed.
- PARCC offers a wide range of accommodations for SWDs (e.g., assistive technology, screen reader, Braille note-taker, extended time, etc.) and ELs (e.g., word-to-word dictionary, speech-to-text for mathematics, general directions provided in a student's native language, etc.).
- Reviewers found the accommodations offered by PARCC to be valid and appropriate based on current research.
- See HUMRRO's report for more: https://www.humrro.org/corpsite/press-release/next-generation-high-school-assessments

Key PARCC Takeaways

PARCC was only one of two tests that is an EXCELLENT or GOOD MATCH to the CCSSO criteria for both ELA and math, in terms of content & depth:

- CONTENT: the test strongly emphasizes the most important content for college and career readiness (CCR) – as called for by CCSS and other CCR standards
- DEPTH: the test requires all students to demonstrate a range of thinking skills, including higher-order skills

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
- Improving balance of literary & informational texts
- Developing the capacity to assess speaking and listening skills
- Addition of more items that assess standards at DOK 1

Areas for Improvement Mathematics

- Increased attention to accuracy of the items—primarily editorial, but in some instances mathematical
- Addition of more items that assess standards at DOK 1 (grade 8)

PARCC changes for 2015-16 tests (driven by member state feedback)

- Consolidated two testing windows into one
- Shortened overall testing time by 1 ½ hours
- Reduced the number of testing units (now includes three units in ELA and three or four in mathematics)

- Also currently considering feasibility of possible item bank

Closing Thoughts

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

- Comparability
- Testing time
- Cost
- Autonomy
- Transparency
- Educator involvement in development

Thank you for your time.

Questions?

vsears@edexcellence.net

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE