


# Evaluating the Content Quality of Assessments

Thanos Patelis  
Center for Assessment


Session at the Center for Assessment's Reidy Interactive Lecture Series (RILS)  
Portsmouth, NH

September 29, 2016



## Overview


- Questions for you
- Conceptual overview of assessment quality
- Overview of criteria and methodology involving in evaluating test content
- See example
- Summary of considerations when implementing



Evaluating Content Quality 2

## Questions


- Do you think that you can implement this in your district or school?
- What are some barriers in your ability to implement this?
- Do you think that if teachers go through this process that they will learn more about what the tests cover?
- Would you *believe* the results of this?



Evaluating Content Quality 3

## Assessment Quality – A Broader Context


- The idea and need for evaluating the quality of tests and testing programs is not new.
  - Oscar Buros, publishing the first *Mental Measurements Yearbook* in 1938, believed it was the responsibility of the profession to monitor itself. He hoped that reviews of tests would influence the quality of tests and test-related materials and research (Carlson & Geisinger, 2012).
- Testing organizations have developed sets of procedures for reviewing quality.
  - For example, for over 30 years, ETS implements internal audits of tests, products and services as part of its ongoing business practice, and utilizes the *ETS Standards for Quality and Fairness* (ETS, 2015) (Wendler, 2015).
  - Additionally, Congress mandated the evaluation of NAEP (Buckendahl, Plake, & Davis, 2009; USDOE, 2009).
- Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.
  - Standards for Educational and Psychological Testing (AERA et al., 2014)
  - Operational Best Practices for Statewide Large-Scale Assessment Programs (CCSSO & A.T.P., 2010)
  - Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2003)
  - Criteria for High-Quality Assessment (Darling-Hammond, Herman, Pellegrino, et al., 2013)
  - CCSSO Criteria for Procuring and Evaluating High-Quality Assessments (CCSSO, 2013)
- Professional organizations have also produced standards that focus on tests for specific purposes (Buckendahl & Plake, 2006).
- Since the beginning of formalized testing, there has been the emphasis to provide the information (technical and non-technical) to the test user to permit the selection and use of quality tests (e.g., Ruch, 1925).
- Further, there has been an outcry to provide independent evaluations of tests for the purposes of informing test users and the general public of the quality of tests (Madaus, 1992).
- ESSA also offers opportunity (and some funding) to evaluate the quality of assessments.



Evaluating Content Quality 4

## A specific effort....


- *CCSSO Criteria for Procuring and Evaluating High Quality Assessments.*
- The Center for Assessment developed methodology for applying these criteria to assessments.



Evaluating Content Quality 5

## CCSSO Criteria

Test Content	Test Characteristics
<b>A. Meet Overall Assessment Goals and Ensure Technical Quality</b>	<b>A. Meet Overall Assessment Goals and Ensure Technical Quality</b>
A.5. Providing accessibility to all students	A.1 Indicating progress toward college and career readiness
A.6. Ensuring transparency of test design and expectations	A.2 Ensuring that assessments are valid for [“single”] purposes
<b>B. Align to Standards – English Language Arts/Literacy</b>	A.3 Ensuring the assessments are reliable
B.1 Assessing student reading and writing achievement	A.4 Ensuring the assessments [“single”] yield valid and consistent score interpretations within and across years
B.2 Focusing on complexity of texts	A.5. Providing accessibility to all students
B.3 Requiring student to read closely and use evidence from texts	A.6. Ensuring transparency of test design and expectations
B.4 Requiring a range of cognitive demand	A.7 Meeting all requirement for data privacy and ownership
<b>B.5. Assessing writing</b>	<b>D. Yield Valuable Reports on Student Progress and Performance</b>
B.6 Emphasizing vocabulary and language skills	D.1 Focusing on student achievement and progress to readiness
B.7 Assessing research and inquiry	D.2. Providing timely data that inform instruction
B.8 Assessing speaking and listening	<b>E. Adhere to Best Practices in Test Administration</b>
B.9 Ensuring high-quality items and a variety of item types	E.1 Maintaining necessary standardization and ensuring test security
<b>C. Align to Standards - Mathematics</b>	
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	



Evaluating Content Quality 6

## CCSSO Criteria - Content

Test Content
<b>A. Meet Overall Assessment Goals and Ensure Technical Quality</b>
A.5. Providing accessibility to all students
A.6. Ensuring transparency of test design and expectations
<b>B. Align to Standards – English Language Arts/Literacy</b>
B.1. Assessing student reading and writing achievement
B.2. Focusing on complexity of texts
B.3. Requiring student 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
B.9. Ensuring high-quality items and a variety of item types
<b>C. Align to Standards - Mathematics</b>
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

- From the 16 CCSSO criteria representing test content, 59 sub-criteria were developed to represent specific features as defined by the criteria.
- Multiple sources of evidence were identified to represent the features indicated in the sub-criteria.
  - Test forms/events
  - Metadata for items and passages
  - Exemplars of accommodations/access features
  - Assessment documentation
- Multiple evaluators would review and evaluate evidence.
  - Developed a process to balance low inference coding and professional judgment

Evaluating Content Quality
7

## Drilling Down: CCSSO Criteria for ELA

**B.1 Assessing student reading and writing achievement in both ELA and literacy:** The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.

**B.2 Focusing on complexity of texts:** The assessments require appropriate levels of text complexity; they raise the bar for text complexity each year so students are ready for the demands of college- and career-level reading no later than the end of high school. Multiple forms of authentic, previously published texts are assessed, including written, audio, visual, and graphic, as technology and assessment constraints permit.

**B.3 Requiring students to read closely and use evidence from texts:** Reading assessments consist of test questions or tasks, as appropriate, that demand that students read carefully and deeply and use specific evidence from increasingly complex texts to obtain and defend correct responses.

**B.5 Assessing writing:** Assessments 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.

Evaluating Content Quality
8

## Drilling Down: Evidences for Each Criterion

Criteria	Evidence
<p><b>B.1</b> Assessing student reading and writing achievement in both ELA and literacy: The assessments are English language arts and literacy tests that are based on an aligned balance of high-quality literary and informational texts.</p>	<ul style="list-style-type: none"> <li>• <b>Test blueprints and other specifications</b> as well as <b>exemplar literary and informational passages</b> are provided for each grade level, demonstrating the expectations below are met.</li> <li>• Texts are balanced across literary and informational text types and across genres, with more informational than literary texts used as the assessments move up in the grade bands, as the state's standards require.</li> </ul> <p><i>For example, for common core aligned assessments, goals include</i></p> <ul style="list-style-type: none"> <li>o In grades 3-8, approximately half of the texts are literature and half are informational;</li> <li>o In high school, because comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational; and</li> <li>o In all grades, informational texts are primarily expository rather than narrative in structure, and in grades 6-12, informational texts are approximately one-third each literary/nonfiction, history/social science, and science/technical.</li> </ul> <ul style="list-style-type: none"> <li>• Texts and other stimuli (e.g., audio, visual, graphic) are previously published or of publishable quality. They are content-rich, exhibit exceptional craft and thought, and/or provide useful information.</li> <li>• History/social studies and science/technical texts, specifically, reflect the quality of writing that is produced by authorities in the particular academic discipline.</li> </ul>

Evaluating Content Quality
9

## Example detailed CCSSO specification

**Type of evidence:** Outcome evidence is derived from examining all Assessment Items. Generalizability evidence from the program documentation.

**Evidence Descriptors:** Description of the characteristics of the sub-criteria, and guidelines of other integrable evidence.

CCSSO Criterion and description of what is to be assessed	Sub-criteria of the CCSSO Criterion to be assessed	Evidence
B.1.1	B.1.1.1	...
B.1.1	B.1.1.2	...
B.1.1	B.1.1.3	...
B.1.1	B.1.1.4	...
B.1.1	B.1.1.5	...
B.1.1	B.1.1.6	...
B.1.1	B.1.1.7	...
B.1.1	B.1.1.8	...
B.1.1	B.1.1.9	...
B.1.1	B.1.1.10	...
B.1.1	B.1.1.11	...
B.1.1	B.1.1.12	...
B.1.1	B.1.1.13	...
B.1.1	B.1.1.14	...
B.1.1	B.1.1.15	...
B.1.1	B.1.1.16	...
B.1.1	B.1.1.17	...
B.1.1	B.1.1.18	...
B.1.1	B.1.1.19	...
B.1.1	B.1.1.20	...
B.1.1	B.1.1.21	...
B.1.1	B.1.1.22	...
B.1.1	B.1.1.23	...
B.1.1	B.1.1.24	...
B.1.1	B.1.1.25	...
B.1.1	B.1.1.26	...
B.1.1	B.1.1.27	...
B.1.1	B.1.1.28	...
B.1.1	B.1.1.29	...
B.1.1	B.1.1.30	...
B.1.1	B.1.1.31	...
B.1.1	B.1.1.32	...
B.1.1	B.1.1.33	...
B.1.1	B.1.1.34	...
B.1.1	B.1.1.35	...
B.1.1	B.1.1.36	...
B.1.1	B.1.1.37	...
B.1.1	B.1.1.38	...
B.1.1	B.1.1.39	...
B.1.1	B.1.1.40	...
B.1.1	B.1.1.41	...
B.1.1	B.1.1.42	...
B.1.1	B.1.1.43	...
B.1.1	B.1.1.44	...
B.1.1	B.1.1.45	...
B.1.1	B.1.1.46	...
B.1.1	B.1.1.47	...
B.1.1	B.1.1.48	...
B.1.1	B.1.1.49	...
B.1.1	B.1.1.50	...

**Identifies the evidence that is provided by the assessment program and assessed by students; the evidence that is provided by the program in instances of coding, and other sources are automatically calculated based on the assessment coding.**

**For B.1.1, the assessment program provides the test forms and materials regarding the test passages. The evidence types indicate the test passages in information tests. The percentage of test passages that are informational is automatically calculated based on the coding system.**

**These coding guidelines provide a rubric for students to score responses.**

**For B.1.1, test items made to have approximately half of the test be informational texts in order to achieve a 50% score.**

**Evidence is derived as possible appropriate comments as well.**

Evaluating Content Quality
10

## Multiple Sources of Evidence

- There are two types of evidence:
  1. Evidence of what the assessment program did (“Outcomes”) – this evidence represented at least two forms of operational items, passages, scoring guides, etc.
  2. Evidence of what the assessment program intended to do (“Generalizability”) – this evidence represented documentation identified by the program (e.g., test specifications)

Evaluating Content Quality
11

## Multiple Evaluators

- Joint panels will evaluate the sub-criteria associated with generalizability.
- Another joint panel will evaluate the programs’ documentation related to accessibility both individually and discuss as a group.
- Evaluators will evaluate the sub-criteria associated with outcomes individually and independently.
  - Produces evaluation rating and comments (rationale, explanation to self, details to remember for group discussion, note to assessment program, etc.)
- Panels of evaluators then discuss as a group and produce a group rating and comment(s) using ratings for both generalizability and outcomes.
  - Ratings are based on evidence
  - Does not have to be consensus—can note majority rating and in Comments disagreements can be noted (indicating what and why)

Evaluating Content Quality
12

### High-Level Overview of Process for Test Content

- For each criterion, CCSSO specified multiple sub-criteria.
- Various training sessions provided to evaluators.
- Evaluators review live forms in format used for actual administration.
- Individual evaluators make a judgment about the evidence provided for particular sub-criteria.
- In sets of facilitated meetings, the evaluators discuss their ratings and the evidence and decide on a group rating.
  - Group level ratings at the criterion-level (not sub-criterion level) are produced which reflect the degree to which the evidence reviewed matches the requirements associated with that criterion – using ratings of “Weak”, “Limited”, “Good”, and “Excellent” with associated scoring guidelines
  - The scores from all the sub-criteria (i.e., outcome, generalizability) are used to inform the criteria-level ratings of the assessment
- Scores of criteria are aggregated to form a score for clusters of criteria labeled “Content” and “Depth”.

Center for Assessment | Evaluating Content Quality | 13

### Differences from Conventional Alignment Methodology

- Considers aspects of test quality in addition to content
  - Appropriateness for all students, including students with disabilities and English learners
- Developed specifically for college/career-ready content standards and assessments
  - Key college/career-ready features evaluated explicitly
- Has explicit evaluation criteria
  - Many other alignment methodologies are descriptive, but not evaluative in the sense of setting an explicit “good enough” criterion and process for applying the criteria to evaluate assessments
- Balances expert judgment with highly specified empirical data
- Considers extensive documentation provided by the assessment developer/publisher in addition to assessment items and content standards—allows evaluation of design and rationale, as well as output
  - This helped support improved documentation
- Produces information useful for general audiences and assessment developers—a combination of high level, easy to understand rating profiles and more detailed ratings and comments

Center for Assessment | Evaluating Content Quality | 14

### Implementation Considerations

- Gather forms, test materials, and metadata
- Select subject matter experts in the content discipline, but they must have experience with individuals being assessed
- Train on using the materials and process.
- Let reviewers take test in mode administered.
- Use coding and scoring sheets.
- Ensure time for discussion.
- Facilitator must ensure engagement and discussion. Moderate discussion to ensure proper communication.
- Do not need consensus; majority is needed; document other comments.
- Make ratings at criterion level and provide rationale.

Center for Assessment | Evaluating Content Quality | 15

### Why?

Validity evidence:

- Certainly, evaluating whether the content expected on an assessment is being represented is important
  - Not just yes/no

Learning:

- Undertaking this effort helps understand what the assessment actually represents as we are looking for whether certain content expectations are being examined.

Center for Assessment | Evaluating Content Quality | 16

### English Language Arts/Literacy

Sub-Criteria	Type	Sub-Criteria	Type
<b>Criterion B.1 (Depth)</b>		<b>Criterion B.4 (Content)</b>	
B.1.1 Informational and literary text balance	Outcome	B.4.1 Vocabulary using tier 2 words, require use of text, and important to central ideas	Outcome
B.1.2 Text quality	Outcome	B.4.2 Writer read/wrote/edited; focus on concrete points, and emphasis; conventions	Outcome
B.1.3 Type of informational texts	Outcome	B.4.3 Percentage of score points devoted to assessing vocabulary	Outcome
B.1.4 Specification of informational and literary balance	Generalizability	B.4.4 Percentage of score points devoted to assessing language	Outcome
B.1.5 Specification of quality of texts	Generalizability	B.4.5 Specifications for vocabulary for college and career readiness	Generalizability
B.1.6 Specification of type of informational texts	Generalizability	B.4.6 Specifications of points for vocabulary	Generalizability
<b>Criterion B.2 (Depth)</b>		B.4.7 Specification of distribution of vocabulary	Generalizability
B.2.1 Justification of texts based on data and qualitative measures of complexity	Outcome	B.4.8 Specifications place sufficient emphasis on vocabulary	Generalizability
B.2.2 Procedures and rationale for how text complexity is measured	Generalizability	<b>Criterion B.7 (Content)</b>	
B.2.3 Documentation specifies target text complexity	Generalizability	B.7.1 Percentage of research skills items requiring analysis, synthesis, and/or organization of info	Outcome
<b>Criterion B.3 (Content)</b>		B.7.2 Significance of research	Outcome
B.3.1 Close reading	Outcome	B.7.3 Specifications on real/simulated research tasks	Generalizability
B.3.2 Central ideas and important particulars	Outcome	<b>Criterion B.8 (Content)</b>	
B.3.3 Questions text dependent and/or open ended	Outcome	B.8.1 Items based on listening skills	Outcome
B.3.4 Questions require direct textual evidence	Outcome	B.8.2 Items based on speaking skills	Outcome
B.3.5 Specification on text dependency	Generalizability	B.8.3 Specifications on listening skills	Generalizability
B.3.6 Specification on proportion of scores devoted to textual evidence	Generalizability	B.8.4 Specification on speaking skills	Generalizability
<b>Criterion B.4 (Depth)</b>		<b>Criterion B.9 (Depth)</b>	
B.4.1 Level of cognitive demand	Outcome	B.9.1 Kinds of formats used on operational forms	Outcome
B.4.2 Procedures for evaluating cognitive demand	Generalizability	B.9.2 Quality of items	Outcome
<b>Criterion B.5 (Content)</b>		B.9.3 Specifications on distribution of item types	Generalizability
B.5.1 Percentages of writing type	Outcome	B.9.4 Alignment to standards & editorial accuracy	Generalizability
B.5.2 Percentages of prompts requiring writing to sources	Outcome		
B.5.3 Specifications of distribution of writing task types	Generalizability		
B.5.4 Specifications require conformation with texts/stimuli	Generalizability		

Center for Assessment | CCSSO Criteria Evaluation Methodology | 17


### Mathematics

Criteria & Sub-Criteria	Type
<b>Criterion C.1 Focusing strongly on the content most needed for success in later mathematics</b>	
C.1.1 Most important content assessed	Outcome
C.1.2 Assessment design reflect important content	Generalizability
C.1.3 The assessment design reflects the standards and reflects a coherent progression of mathematics content from grade to grade and course to course	Generalizability
<b>Criterion C.2 Assessing a balance of concepts, procedures, and applications</b>	
C.2.1 Balance of % of points conceptual understanding, procedural skills and fluency, & applications	Outcome
C.2.2 Balance of conceptual understanding, procedural skills and fluency, & applications	Generalizability
C.2.3 Specifications on all math categories for students at all performance levels	Generalizability
Assesses the depth that reflect the demands of College and Career Readiness (Cluster)	
<b>Criterion C.3 Connecting practice to content</b>	
C.3.1 Meaningful connections between practices and content	Outcome
C.3.2 Specifications & explanation of assessing math practices with content	Generalizability
C.3.3 Assessments for each grade and course meaningfully connect mathematical practices and processes with mathematical content (especially with the most important mathematical content at each grade)	Generalizability
<b>Criterion C.4 Requiring a range of cognitive demand</b>	
C.4.1 Cognitive Demand	Outcome
C.4.2 Specification of Cognitive Demand	Generalizability
<b>Criterion C.5 Ensuring high-quality items and a variety of item types</b>	
C.5.1 Distribution of item types	Outcome
C.5.2 Degree of high-quality items	Outcome
C.5.3 Specification of item types and quality	Generalizability
C.5.4 Specification of distribution of item types	Generalizability

Center for Assessment | CCSSO Criteria Evaluation Methodology | 18

## Scoring

- Each **outcome** sub-criterion is scored as 0, 1, or 2 for each assessment form and results for all outcome sub-criteria are added together to form a criterion score. (Scoring rules are provided.)
  - Depending on the number of sub-criteria (specific scoring rules are offered), the criterion score is classified as:
    - Excellent Match
    - Good Match
    - Limited Match
    - Weak Match
- Each **generalizability** sub-criterion is scored as 0, 1, or 2 for each assessment overall and results are added together for a criterion score. (Scoring rules are provided.)
  - The larger score here can be used to increase the classification for the criterion
  - A moderate score can be used to support/keep the classification for the criterion
  - A smaller score here can be used to decrease the classification for the criterion



Evaluating Content Quality

19

## Reporting Approach – Aligned to CCSSO Criteria

Results of Applying the CCSSO Criteria for High-Quality Assessments


	Degree of Match to CCSSO Criteria			
	None Match	Limited Match	Good Match	Excellent Match
<b>A. Meet Overall Assessment Goals and Technical Quality</b>				
• A.5: Providing accessibility to all students, including English learners and students with disabilities (subset of the criterion)	●	●	●	●
• A.6: Ensuring transparency of test design and expectations	●	●	●	●
<b>B. English Language Arts/Literacy</b>				
<b>1. Assesses the <i>content</i> most needed for College and Career Readiness</b>				
([Summary of rationale and other comments])				
• B.3: Requiring students to read closely and use evidence from texts	●	●	●	●
• B.5: Assessing writing	●	●	●	●
• B.6: Emphasizing vocabulary and language skills	●	●	●	●
• B.7: Assessing research and inquiry	●	●	●	●
• B.8: Assessing speaking and listening	●	●	●	●
<b>2. Assesses the <i>depth</i> that reflect the demands of College and Career Readiness</b>				
([Summary of rationale and other comments])				
• B.1: Assessing student reading and writing achievement in both ELA and literacy	●	●	●	●
• B.2: Focusing on complexity of texts	●	●	●	●
• B.4: Requiring a range of cognitive demand	●	●	●	●
• B.9: Ensuring high-quality items and a variety of item types	●	●	●	●

Four ordered categories with the second two representing acceptable levels.

Results will be organized by each criterion, but scores will be provided by each sub-criterion.

Two aggregated scores will be provided. The level of aggregation represents the set of criteria associated with "content" and "depth".

Comments will be provided with rationale and feedback.



CCSSO Criteria Evaluation Methodology

20

## References

AERA, APA, NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA, Buckendahl, C. W. & Plake, B. S. (2006). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 725-738). Mahwah, NJ: Lawrence Erlbaum Associates.

Buckendahl, C. W., Plake, B. S., & Davis, S. L. (2009). Conducting a lifecycle audit of the National Assessment of Educational Progress. *Applied Measurement in Education*, 22, 321-338.

Carlson, J. F. & Geisinger, K. F. (2012). Test reviewing at the Buros Center for Testing. *International Journal of Testing*, 12, 122-135.

Council of Chief State School Officers & Association of Test Publishers (2010). *Operational best practices for statewide large-scale assessment programs*. Washington, DC: Council of Chief State School Officers. (ISBN: 978-1453716281)

Council of Chief State School Officers (2013). *Criteria for procuring and evaluating high-quality assessments*. Washington, DC: Author. <http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf>

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

Educational Testing Service (2015). *ETS Standards for Quality and Fairness*. Princeton, NJ: Author.


Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Madaus, G. F. (1992). An independent auditing mechanism for testing. *Educational Issues: Measurement and Practice*, 11(1), 26-31. DOI: 10.1111/j.1745-3992.1992.tb00225.x

Ruch, G. M. (1925). Minimum essentials in reporting data on standard tests. *Journal of Educational Research*, 12, 349-358.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Evaluation of the National Assessment of Educational Progress, Study Reports, Washington, D.C., 2009.

Wendler, C. (April, 2015). Ensuring assessment quality: the ETS internal audit process. Presentation at the annual meeting of the National Council on Measurement in Education, Chicago, IL.




Evaluating Content Quality


21

## For more information:

Center for Assessment  
[www.nciea.org](http://www.nciea.org)



Brian Gong – [brong@nciea.org](mailto:brong@nciea.org)  
 Erika Hall – [ehall@nciea.org](mailto:ehall@nciea.org)  
 Susan Lyons – [slyons@nciea.org](mailto:slyons@nciea.org)  
 Thanos Patelis – [tpatelis@nciea.org](mailto:tpatelis@nciea.org)



Evaluating Content Quality

22