The Center for Assessment developed methodology... Would you... Do you think that if teachers go through this process... Do you think that you can implement this in your... What are some barriers in your ability to implement this?... Do you think that if teachers go through this process... Would you believe the results of this?

A specific effort....

• CCSSO Criteria for Procuring and Evaluating High Quality Assessments.

• The Center for Assessment developed methodology for applying these criteria to assessments.

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

Assessment Quality – A Broader Context

• The idea and need for evaluating the quality of tests and testing programs is not new.

• 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).

• Professional organizations have also produced standards that focus on tests for specific purposes (Buckendahl & Plake, 2006).

• Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.

• Assessment Quality – A Broader Context

• The idea and need for evaluating the quality of tests and testing programs is not new.

• 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).

• Professional organizations have also produced standards that focus on tests for specific purposes (Buckendahl & Plake, 2006).

• Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.

• The idea and need for evaluating the quality of tests and testing programs is not new.

• 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).

• Professional organizations have also produced standards that focus on tests for specific purposes (Buckendahl & Plake, 2006).

• Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.

• The idea and need for evaluating the quality of tests and testing programs is not new.

• 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).

• Professional organizations have also produced standards that focus on tests for specific purposes (Buckendahl & Plake, 2006).

• Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.

CCSSO Criteria

<table>
<thead>
<tr>
<th>Test Content Characteristics</th>
<th>Criteria Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Meet Overall Assessment Goals and Ensure Technical Quality</td>
<td>A. Ensuring the assessments are reliable. B. Ensuring the assessments are valid. C. Ensuring the assessments are fair.</td>
</tr>
<tr>
<td>B. Align to Standards – English Language Arts/Literacy</td>
<td>A. Ensuring the tests align to the standards. B. Ensuring the tests assess the standards effectively.</td>
</tr>
<tr>
<td>C. Aligned to Standards – Mathematics</td>
<td>A. Ensuring the tests align to the standards. B. Ensuring the tests assess the standards effectively.</td>
</tr>
<tr>
<td>D. Yield Valuable Reports on Student Progress and Performance</td>
<td>A. Ensuring the reports are timely. B. Ensuring the reports are useful.</td>
</tr>
<tr>
<td>E. Adhere to Best Practices in Test Administration</td>
<td>A. Ensuring the tests adhere to best practices. B. Ensuring the tests are implemented as intended.</td>
</tr>
</tbody>
</table>

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
Drilling Down: CCSSO Criteria for ELA

8.1 Assessing student reading and writing achievement in both English Language Arts (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.

8.2 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, visual, audio, and graphic, as technology and assessment constraints permit.

8.3 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.

8.5 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

- 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 forms and passages
  - Examples of accommodations/access features
  - Assessment documentation
- Multiple evaluators would review and evaluate evidence.
  - Developed a process to balance low (inferencing and professional judgment)
  - 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 states’ standards require.
  - In high school, because comprehension of complex informational texts is crucial for readiness, texts are approximately one-third literature and two-thirds informational.
  - 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.
  - For example, for common core aligned assessments, goals include:
    - 1. Evidence of what the assessment program did
    - 2. Evidence of what the assessment program intended to do
    - 3. Evidence of what the assessment program is doing
    - 4. Evidence of what the assessment program might do
    - 5. Evidence of what the assessment program should do
    - 6. Evidence of what the assessment program will do
    - 7. Evidence of what the assessment program has done
    - 8. Evidence of what the assessment program will have done
    - 9. Evidence of what the assessment program will have been done
    - 10. Evidence of what the assessment program will have not been done
    - 11. Evidence of what the assessment program will not have done
    - 12. Evidence of what the assessment program would have done
    - 13. Evidence of what the assessment program would have not done
    - 14. Evidence of what the assessment program should have done
    - 15. Evidence of what the assessment program should have not done
    - 16. Evidence of what the assessment program cannot do

- 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)

- 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)
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 criterion-level ratings of the assessment.
- Scores of criteria are aggregated to form a score for clusters of criteria labeled “Content” and “Depth”.

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
  - In college/career-ready features evaluated explicitly
- Has explicit evaluation criteria
  - Many other alignment methodologies are descriptive, but not evaluation 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

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 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.

English Language Arts/Literacy

<table>
<thead>
<tr>
<th>Sub-Criteria Type</th>
<th>Criterion</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.3</td>
<td>Type</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.1.4</td>
<td>Type</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.1.6</td>
<td>Type</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.2.2</td>
<td>Procedures and rationale for how text complexity is measured</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.3.5</td>
<td>Specification on text-dependency</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.3.6</td>
<td>Specification on proportion of scores devoted to textual generalizability</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.4.2</td>
<td>Procedures for evaluating cognitive demand</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.5.1</td>
<td>Percentages of writing type</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.5.2</td>
<td>Percentages of prompts requiring writing to sources</td>
<td>Outcome</td>
</tr>
<tr>
<td>B.5.3</td>
<td>Specification of distribution of writing tasks/types</td>
<td>Outcome</td>
</tr>
</tbody>
</table>

Mathematics

<table>
<thead>
<tr>
<th>Sub-Criteria Type</th>
<th>Criterion</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1.1</td>
<td>Most important content assessed</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.1.2</td>
<td>Most important content assessed</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.1.3</td>
<td>The assessment design reflects important content</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.2.1</td>
<td>Balance of % of points conceptual understanding, procedures, and applications</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.2.2</td>
<td>Balance of % of points conceptual understanding, procedures, and applications</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.2.3</td>
<td>Specifications on all math categories for students at all performance levels</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.3.1</td>
<td>Meaningful connections between practices and content</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.3.2</td>
<td>Specifications &amp; explanation of assessing math practices with content</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.4.2</td>
<td>Specification of Cognitive Demand</td>
<td>Outcome</td>
</tr>
<tr>
<td>C.5.1</td>
<td>Specification of Cognitive Demand</td>
<td>Outcome</td>
</tr>
</tbody>
</table>

9/29/2016
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

For more information:

Center for Assessment
www.nciea.org

Brian Gong – bgong@nciea.org
Erika Hall – ehall@nciea.org
Susan Lyons – slyons@nciea.org
Thanos Patelis – tpatelis@nciea.org

References