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



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

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 liself.
 He hoped that reviews of tests would influence the quality of tests and test-related materials and research (Carlson & Geisinger, 2012).
- - For example, for over 30 years, ETS implements internal audits of tests, products and services as part of its ongoing bus utilizes the ETS Standards for Quality and Fairness (ETS, 2015) (Wendler, 2015).
- Standards, guidelines, and best practices have been published to provide test developers, publishers, and users with expectations, appropriate practices, and important characteristics of tests.
- with expectations, appropriate practices, and important characteristics of tests.

 Standards for Educational and Psychological Testing (AEAR et al., 2014)

 Operational Best Practices for Statewised Large-Scale Assussment Programs (CCSSO & ATP, 2010)

 Code of Far Testing Practices in Education (Joint Committee on Testing Practices, 2003)

 Citatria for light Quality Assussment (Joint Patemond, Herman Pelignijoc, 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.



A specific effort....

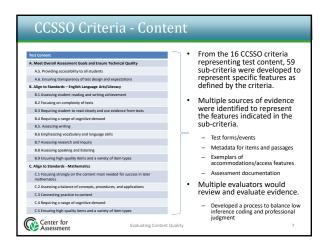
- CCSSO Criteria for Procuring and Evaluating High Quality Assessments.
- The Center for Assessment developed methodology for applying these criteria to assessments.

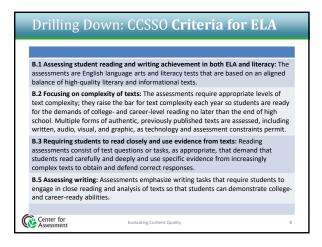


Evaluating Content Quality

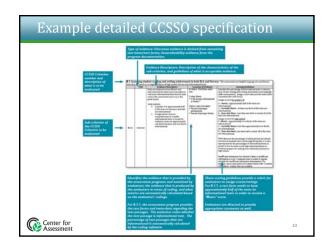
CCSSO Criteria A.4 Ensuring the assessments ["snip"] yield valid and co interpretations within and across years B.2 Focusing on complexity of texts
B.3 Requiring student to read closely and use evidence from texts A.6. Ensuring transparency of test design and expectations B.4 Requiring a range of cognitive demand B.5. Assessing writing
B.6. Emphasizing vocabulary and language skills
B.7 Assessing research and inquiry D. Yield Valuable Reports on Student Progress and Performance
D.1 Focusing on student achievement and progress to readines
D.2. Providing timely data that inform instruction E. Adhere to Best Practices in Test Administration
E.1 Maintaining necessary standardization and ensuring test : B.8 Assessing speaking and listening B.9 Ensuring high-quality items and a variety of item types

Align to Standards - Mathematics C.1 Focusing strongly on the content most needed for success in later mathematics C.3 Connecting practice to content C.5 Ensuring high-quality items and a variety of item types Evaluating Content Quality









Multiple Sources of Evidence

- There are two types of evidence:
 - Evidence of what the assessment program did ("Outcomes") – this evidence represented at least two forms of operational items, passages, scoring guides, etc.
 - 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

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
 - 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
- Ratings are based on evidence

individually and independently.

 Does not have to be consensus—can note majority rating and in Comments disagreements can be noted (indicating what and why)



Evaluating Content Quality

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".



Control of Control Control

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 learner.
- · 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



Evaluating Content Quality

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.
- $\bullet \quad \hbox{Do not need consensus; majority is needed; document other comments.}\\$
- Make ratings at criterion level and provide rationale.



Evaluating Content Quality

Why?

Validity evidence:

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

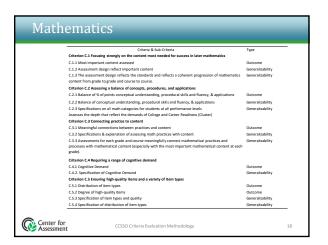
Learning:

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



Evaluating Content Quality

English Language Arts/Literacy B.1.1 informational and literary text balance Outcome Differences B.1.3 repe of informational and literary text balance Differences B.1.3 repe of informational and literary balance Gliconic literature Gliconic li ext balance Outcome Outcome ixts Generalizability rmational texts Generalizability 8.1.5 Specification or types or several control of City City (Digits) 8.2.1 sectification of texts based on data and qualitative measures of complexity Section 1. S reaciness Generalizability B.6.6 Specifications of points for vocabulary Generalizability Criterion 8.3 (Content) 8.3.2 Central ideas and important particulars 8.3.2 Central ideas and important particulars 8.3.3 Central ideas and important particulars 8.3.4 Specification test dependent part upon sees death 1.3.4 Central ideas and important particulars 8.3.5 Specification on report central ideas 8.3.5 Specification on reportion of scores devoted to testing. Listenon B./ (Content) B.7.1 Percentage of research skills items requiring analysis, psychesis. KVcs organization of info B.7.2 Significance of research Generalizabilities e of research Generalizability ons on real/simulated research tasks Generalizability 8.7.3 Specifications contravalments 8.8.1 Step Saud on Section 2.8.1 Contraval 0.8.2 Contraval 0.8.3 Contraval 0.8.3 Contraval 0.8.3 Contraval 0.8.3 Contraval 0.8.3 Specifications on Usering 24% 8.8.3 Specification on speaking 38% 8.8.4 Specification on speaking 38% 6.8.4 Specification on speaking 38% 6.8.4 Specification on speaking 38% 6.8.4 Specification on speaking 38% evidence Generalizability Effetion BA (Bright) Effetion BA (Bright) BA 1.twel of cognitive demand BA 2.twel of cognitive demand BA 3.twel of cognitive demand BA 3.twel of cognitive demand BA 3.twellow of common the cognitive demand BA 3.twellow of common the cognitive demand BA 3.twellow of cognitive demand of cogn - Generalizability ns Outcome Outcome ications on distribution of item types Generalizability nent to standards & editorial accuracy Grind all 2ability



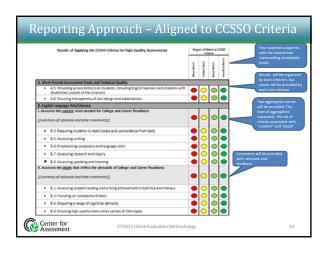
Scoring

- Each <u>outcome</u> 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



References

Evaluating Content Quality



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. Weichendahl, C. W., Plake, B. S., & Davis, S. I. (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 Assessment of Educational Progress. Applied Measurement in Education, 22, 321-338. Council of Chief State School Officers & Association of Test Publishers (2010). Operational best practices for statewide large-selle sessessment programs. Washington, DC: 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%20Assess ment. 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%20Assess ment. School Officers (2013). Esta School Officers (2013). Criteria for Ingh-quality assessment. 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. Measurement and Practice, 11(1), 26-31. DOI: 10.1111/j.1745-3992.1992.tb00225 x. Washington, DC., 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

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 Center for Assessment Evaluating Content Quality