The future is here: The (new) *Standards for Educational and Psychological Testing*

Enis Doğan

Enis.Dogan@ed.gov

Research Fellow/ Resident Expert

at NCES

January, 13 2016

Overview

PART I:

- Background on the *Standards a*nd the revision process
- 1999 versus 2014 standards
  - Validity
  - Reliability/precision
  - Fairness
  - Design and Development
  - Testing in Program Evaluation and Public Policy and Accountability
- Standards summarized in 140 characters or less
- Some practical recommendations
- Take-home self quiz

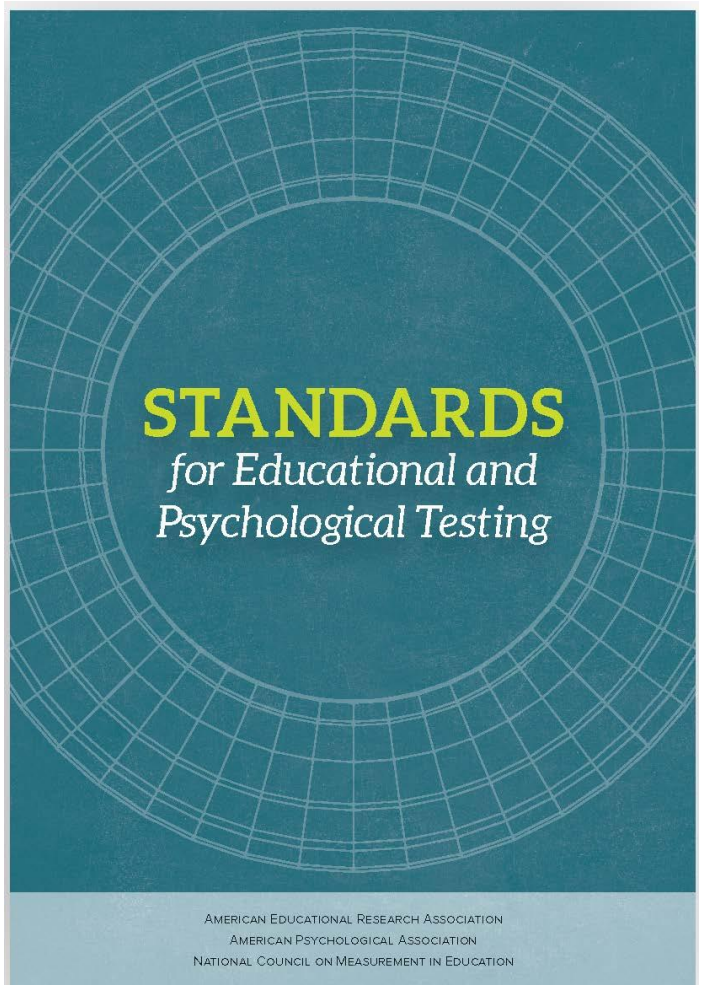
PART II:

- Example of a validity evidence framework based on the *Standards*
- Other helpful resources

PART I

The Standards

- "Definitive technical, professional and operational standards for all forms of assessments that are professionally developed and used in a variety of settings" (Camara, 2014)



The Standards

- 1st edition: *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954)
- 2nd edition, published (jointly by APA, AERA, and NCME) in 1966 and called *Standards for Educational and Psychological Tests and Manuals*,
- 3rd edition published in 1974
- The 1985 edition - titled as *Standards for Educational and Psychological Testing*- represented a shift toward a unitary concept in validity theory,
- In 1999, the *Standards* were again revised,
  - highlighted that validity and reliability were functions of the interpretations of test scores for their intended uses and not of the test itself.

Revision of the 1999 Standards

- January 2009: First Meeting of the Joint Committee
- January 2011: Initial Draft
- January 2011-April 2011: Public Review
- May 2011-March 2012: Joint Committee reviewed and responded comments
- Spring of 2012-Fall 2014: Revised Standards sent to AERA, NCME, and APA for final review

1999 versus 2014

1999 Standards

Introduction

Part I: Test Construction, Evaluation, and Documentation
- Validity
- Reliability and Errors of Measurement
- Test Development and Revision
- Scales, Norms, and Score Comparability
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

Part II: Fairness in Testing
- Fairness in Testing and Test Use
- The Rights and Responsibilities of Test Takers

- Testing Individuals of Diverse Linguistic Backgrounds
- Testing Individuals with Disabilities

Part III: Testing Applications
- The Responsibilities of Test Users
- Psychological Testing and Assessment
- Educational Testing and Assessment
- Testing in Employment and Credentialing

- Testing in Program Evaluation and Public Policy

2014 Revised Standards

Introduction

Part I: Foundations
- Validity
- Reliability/Precision and Errors of Measurement
- Fairness in Testing

Part II: Operations
- Test Design and Development
- Scores, Scales, Norms, Score Linking, and Cut Scores

- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests
- Rights and Responsibilities of Test Takers
- Rights and Responsibilities of Test Users

Part III: Testing Applications
- Psychological Testing and Assessment
- Workplace Testing and Credentialing
- Educational Testing and Assessment
- Testing in Program Evaluation and Public Policy and Accountability

The Standards

1999 Standards

Introduction

Part I: Test Construction, Evaluation, and Documentation

- Validity
- Reliability and Errors of Measurement
- Test Development and Revision
- Scales, Norms, and Score Comparability
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

2014 Revised Standards

Introduction

Part I: Foundations

- Validity
- Reliability/**Precision** and Errors of Measurement
- Fairness in Testing

Validity

- "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests"(p.11).

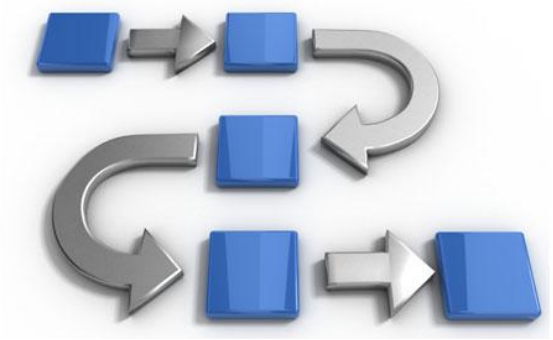
- Validity evidence is required for each and every use (e.g. classifying students, measuring student growth, prediction, accreditation)

Validity

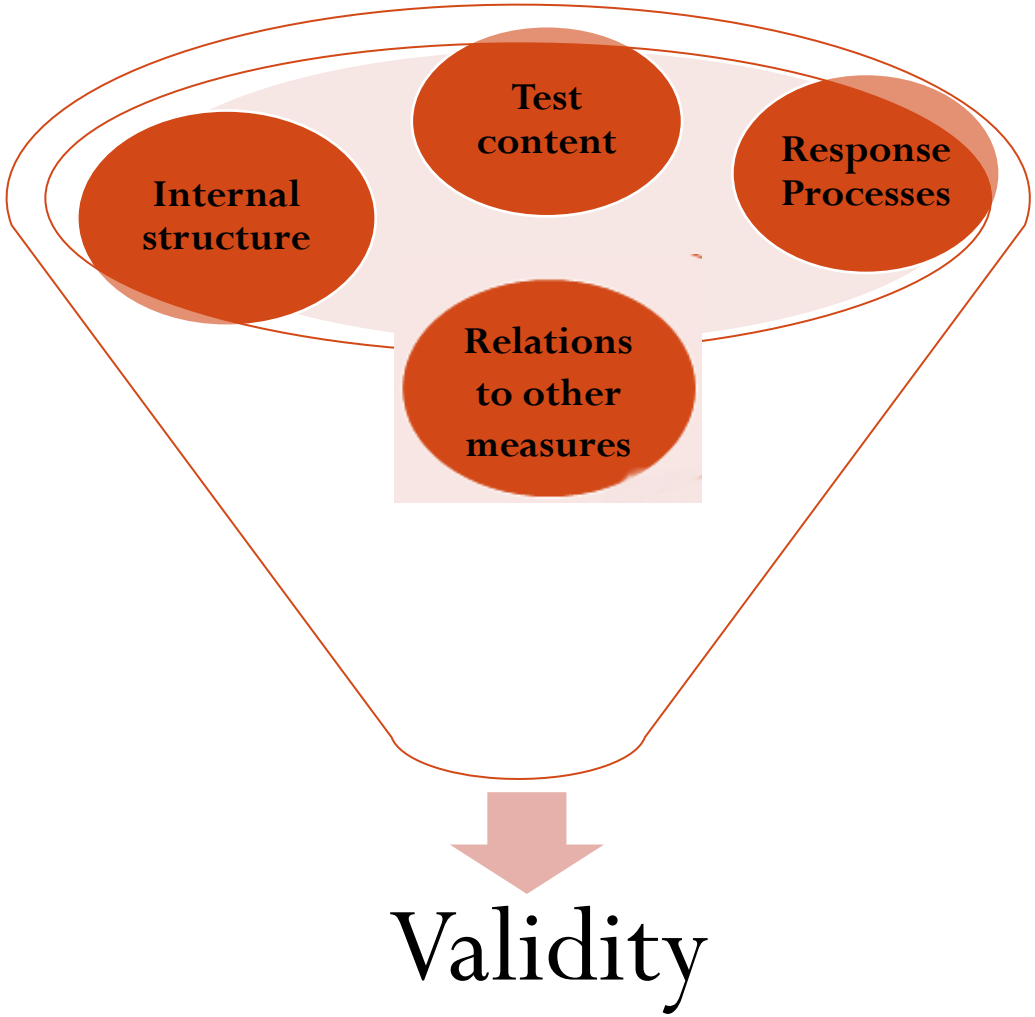
- Validation is not an activity that occurs once the assessments are developed, but rather is an ongoing process (Messick, 1995) that is initiated at the beginning of assessment design and continues throughout development and implementation
- Validity evidence takes one of two forms (Haladyna, 2006):

i. Empirical

ii. Procedural

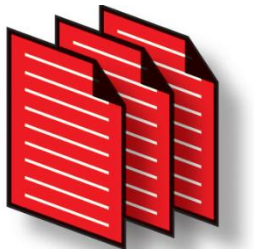


Sources of validity evidence



Validity evidence based on test content

- Major threats: construct *underrepresentation* and construct *contamination*
- Procedural evidence
  - Test specifications including frameworks and blueprints
    - Content and relative importance of aspects of the content
    - Cognitive skills and rigor
  - Bias and sensitivity guidelines and reviews
  - Use of ECD or other principled approach to design and development
- Empirical evidence
  - Alignment studies



Evidence Centered Design

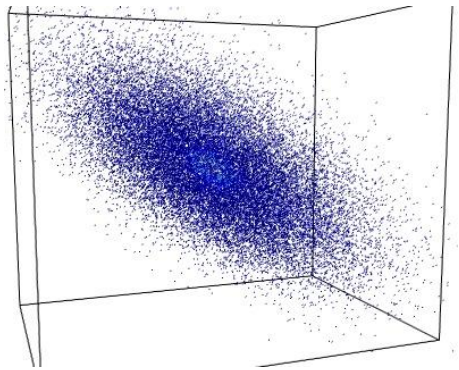
Validity evidence based on response processes

- Major threats: mismatch between intended and actual cognitive processes items evoke; items susceptible to test-taking strategies
- Procedural evidence
  - Definition of cognitive skills and rigor
  - Use of ECD or other principled approach to design and development
  - Clear directions
- Empirical evidence

- **Cognitive lab studies**
- **Analysis of log/process data**

Validity evidence based on internal structure

- Major threats: poorly specified or misspecified dimensionality and structure, items with poor psychometric quality
- Procedural evidence
  - Frameworks and test specifications
  - Form construction specifications
- Empirical evidence
  - Dimensionality studies
  - Differential Item Functioning (DIF) studies



Validity evidence based on relation to other variables

- Major threats: unclear or poorly justified prediction criteria, differential prediction for different subgroups
- Procedural evidence
  - Active involvement of experts and stakeholders in to-be-predicted domain in development of prediction criteria and assessment content
- Empirical evidence
  - Predictive and concurrent studies
  - Judgmental studies involving experts and stakeholders from the to-be-predicted domain

The Standards

1999 Standards

Introduction

Part I: Test Construction, Evaluation, and Documentation

- Validity
- Reliability and Errors of Measurement
- Test Development and Revision
- Scales, Norms, and Score Comparability
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

2014 Revised Standards

Introduction

Part I: Foundations

- Validity
- Reliability/**Precision** and Errors of Measurement
- Fairness in Testing

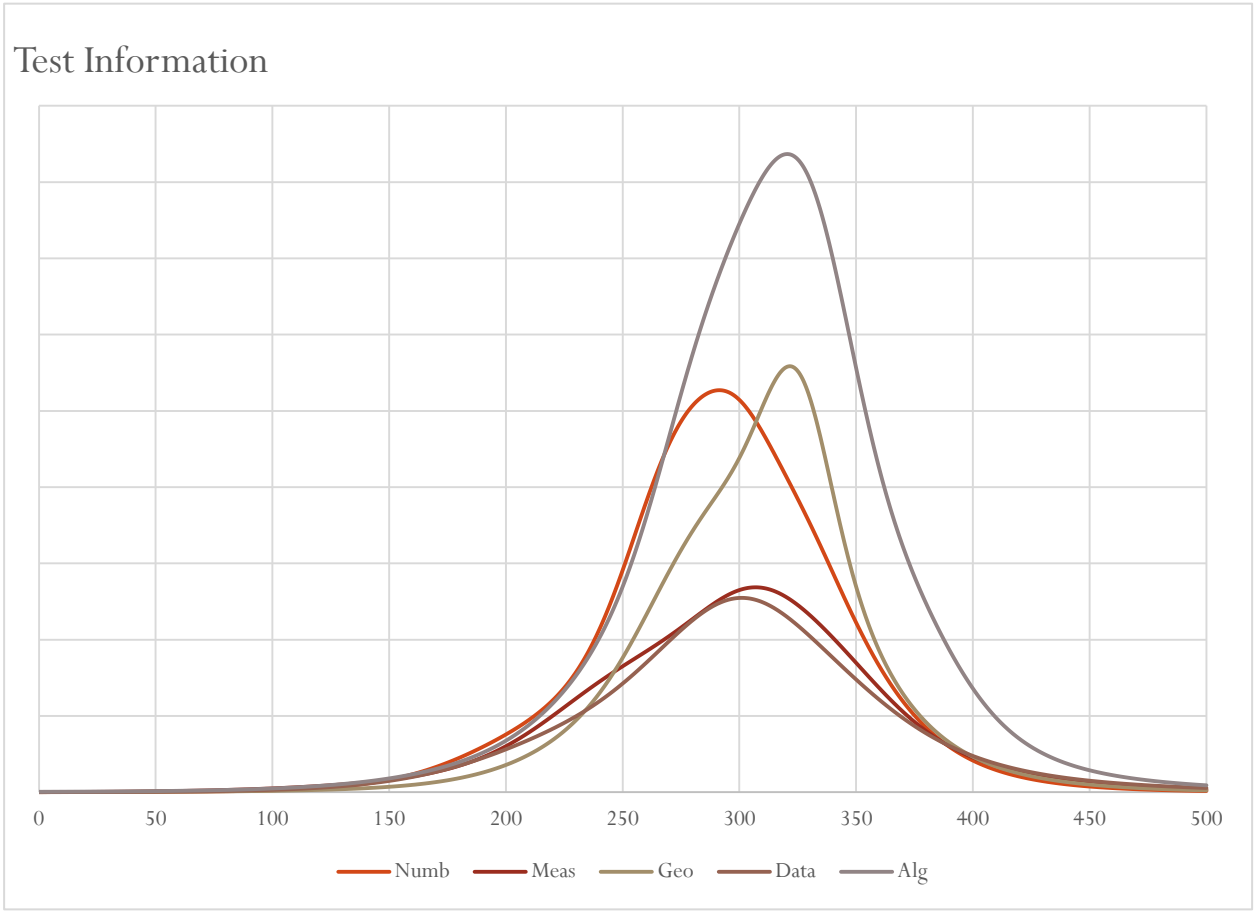
Reliability/Precision and Errors of Measurement

- Greater emphasis on conditional precision over overall reliability
- Call for documenting decision consistency and accuracy
- Questions to consider: Would the students' scores change
  - had they been given a different set of items, or other stimuli
  - had they responses been scored by a different scorer
  - had they been tested at a different point in time
- Generalizability Theory: pinpoints the sources of measurement error, disentangles them, and estimates each one
- IRT provides a powerful tool to deal with reliability/precision: Test Information Functions

Reliability/Precision and Errors of Measurement

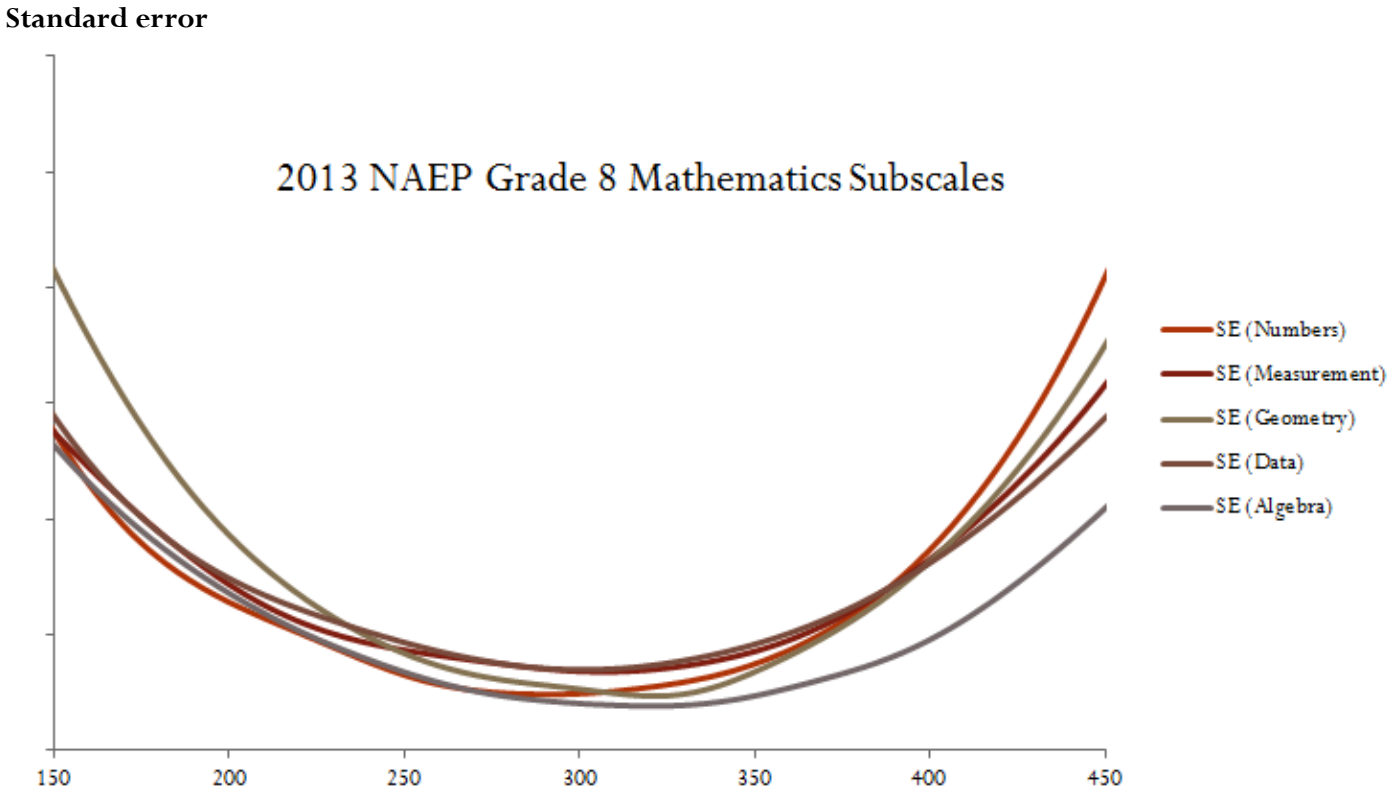
- **Standard 2.3** For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.
- **Standard 2.14** When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported at the vicinity of each cut score.
- **Standard 2.16** When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

Reliability/Precision and Errors of Measurement



Standard 2.3 estimates of reliability/precision for each total score, subscore

Reliability/Precision and Errors of Measurement



Standard 2.14 conditional standard errors of measurement at several score levels

The Standards

1999 Standards

Part II: Fairness in Testing

- Fairness in Testing and Test Use
- The Rights and Responsibilities of Test Takers
- Testing Individuals of Diverse Linguistic Backgrounds
- Testing Individuals with Disabilities

2014 Revised Standards

Part I: Foundations

- Validity
- Reliability/Precision and Errors of Measurement
- Fairness in Testing

Part II: Operations

- Rights and Responsibilities of Test Takers

Fairness in Testing

- "Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use" (p. 49)

- The ultimate question: How do we best enable ALL students to demonstrate what they know and can do.
- Universal Design is your BFF
  - Precisely Defined Constructs
  - Simple, Clear, and Intuitive Instructions and Procedures
  - Maximum Readability and Comprehensibility
  - Maximum Legibility
- Guidelines for accessibility, fairness, accommodations, bias and sensitivity reviews and

Fairness in Testing

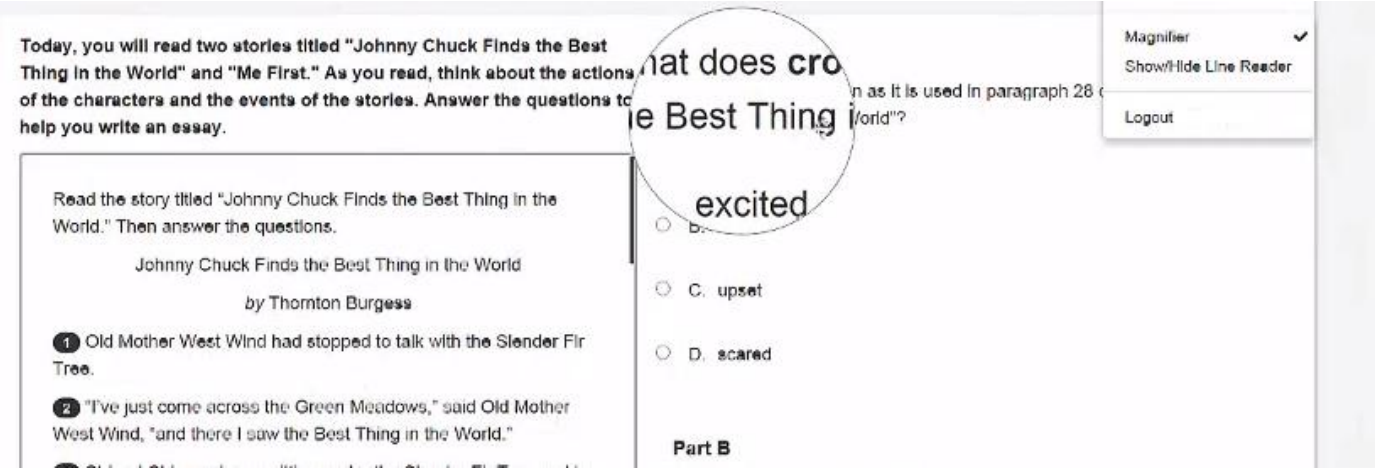
- **Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests being affected by construct-irrelevant such as linguistic, communicative, cognitive, cultural, physical or other characteristics.

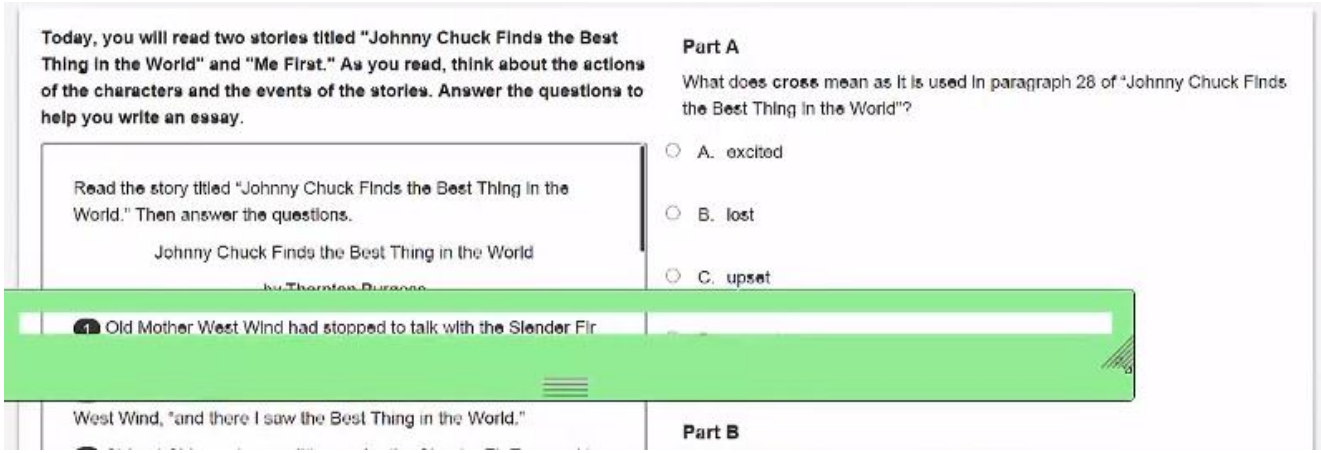
- **Standard 3.3** Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.
- **Standard 3.9** Test developers and/or test users are responsible for developing and providing test accommodations when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

Examples of common accessibility features in digital assessments

- Available to all students
  - audio amplification, eliminate answer choices, flag items for review, highlight tool, magnification/enlargement, pop-up glossary, spell checker and writing tools (e.g. (bold, italic, underline, bulleted text, copy, paste etc.).
- Available based on Personal Needs Profile (PNP)
  - background/font color contrast, general administration directions clarification , line reader tool, masking, and text-to-speech for the mathematics assessments.

Examples of common accessibility features in digital assessments





Examples of common accommodations in digital assessments

- For SWDs
  - read aloud or text-to-speech in the ELA/Literacy assessments (assistive technology, braille edition, closed-captioning of video, tactile graphics, extended time, scribe or speech-to-text (i.e., dictating/ transcription) for the ELA/Literacy assessments, and video of a human interpreter for the ELA/Literacy assessments for students who are deaf or hard of hearing.
- For English Language Learners
  - English/Native language word-to-word dictionary (ELA/Literacy & mathematics), read aloud or text-to-speech in English (ELA/Literacy), scribe or speech-to-text, extended time and frequent breaks.

The Standards

1999 Standards
Introduction

Part I: Test Construction, Evaluation, and Documentation

- Validity
- Reliability and Errors of Measurement
- Test Development and Revision
- Scales, Norms, and Score Comparability
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

2014 Revised Standards
Introduction

Part I: Foundations

- Validity
- Reliability/**Precision** and Errors of Measurement
- Fairness in Testing

Part II: Operations

- Test **Design** and Development
- Scores, Scales, Norms, Score Linking, and Cut Scores
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

Operations: Design and Development

- Emphasis on considering validity, fairness, and precision before test development begins
- Psychometric specifications including
  - Statistical properties of individual items and the whole test (e.g. difficulty, discrimination)
  - Properties of the reporting scale
  - Evaluation of model assumptions and model fit
- Use of technology
  - Scoring specs used by automated scoring engines
  - Item selection and content coverage in adaptive tests
  - Interoperability in systems used for item banking, form assembly, and test administration

Operations: Design and Development

- **Standard 4.0** Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.
- **Standard 4.2** In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. … Specifications for computer-based tests should include a description of any hardware and software requirements.

The Standards

1999 Standards

Part III: Testing Applications

- The Responsibilities of Test Users
- Psychological Testing and Assessment
- Educational Testing and Assessment
- Testing in Employment and Credentialing

- Testing in Program Evaluation and Public Policy

2014 Revised Standards

Part III: Testing Applications

- Psychological Testing and Assessment
- Workplace Testing and Credentialing
- Educational Testing and Assessment
- Testing in Program Evaluation and Public Policy **and Accountability**

Testing in Program Evaluation and Public Policy and Accountability

- Valid interpretations depend on clear description of "how samples were formed and how the tests were designed, scored, and reported", and the extent to which the sample is representativeness of the population the inferences are about.

- Validity evidence for the intended uses of an assessment is not sufficient if the outcomes are to be used for accountability purposes. Such use requires additional validity evidence.

Testing in Program Evaluation and Public Policy and Accountability

- **Standard 13.3** When accountability indices, indicators of effectiveness on program evaluations or policy studies, or other models (such as value-added models) are used, the method for constructing such indices, indicators, or models should be described and justified, and their technical qualities should be reported .
- **Standard 13.4** Evidence of validity, reliability, and fairness for each purpose for which a test is used in a program evaluation, policy study, or accountability system should be collected and made available.
- **Standard 13.6** Reports of group differences in test performance should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of the differences. If appropriate contextual information is not available, users should be cautioned against misinterpretation.

The Standards summarized

- Clearly describe the construct being measured, explain the intended uses of the assessment; and **create and implement specifications and procedures** that would allow users to make valid inferences.

- Specify, implement, check, document, repeat.

Some practical recommendations: Validity

- Be explicit about what is and is not construct relevant
- Build a framework around which validity evidence can be gathered and organized
- Document all procedural and empirical validity evidence in one place.

- Educate users
  - **Carr, P., Dogan, E., Tirre, W., & Walton, E. (2007).** Large-Scale Indicator Assessments: What Every Educational Policymaker Should Know. In Moss, P. A. (Ed.), *Evidence and Decision Making: 2007 National Society for the Study of Education (NSSE) Yearbook*, (pp. 328-347). Oxford: Blackwell.
  - **Mazzeo, J., Lazer, S., & Zieky, M.J. (2006).** Monitoring Educational Progress with Group-Score Assessments. In Brennan, R. L. (Ed.). *Educational measurement* (4th ed.). Westport, CT: ACE/Praeger Publishers.

Some practical recommendations

- **Reliability/precision:**
  - Use your technical report as an organizer/enabler/enforcer
  - Include TIFs and CSEM plots in technical reports
  - Include classification consistency and accuracy estimates in technical reports
- **Fairness**
  - Develop clear and distinct definitions for accessibility features and accommodations
  - Collect validity evidence for interpretations regarding all relevant subgroups
  - Adhere to Universal Design principles

Some practical recommendations

- **Design and development**
  - Think like an engineer
  - Consider ECD
  - Engage psychometricians at development phase
    - New technology-based item types
    - Scoring rules
  - Justify use of new technology -enhanced items
    - Use technology "enhancements" only when you cannot measure the construct without them
    - Ask if the benefit exceeds the risk/cost?

Take home self-quiz

Validity:

1. What kinds of validity evidence does the program collect? Can one easily access the validity evidence collected so far?
2. Where and how are the boundaries of the constructed being measured defined/explained?
3. What is the definition of cognitive complexity and do blueprints indicate the distribution of items according to complexity level?
4. How does the program decide what study (validity research) to pursue/fund?

Reliability/precision

1. Are Test Information Functions and CSEM plots included in the technical reports?
2. Does the program document classification accuracy and consistency?

Fairness

1. What are accessibility features used in the assessment?
2. What are accommodations provided in the assessment?
3. When were the bias and sensitivity guidelines last updated?
4. What subgroups are the Differential Item Functioning (DIF) analyses conducted for?

PART II

A validity evidence framework based on the Standards

- Validation is an ongoing process (Messick, 1995, p. 740) that is initiated at the beginning of assessment design and continues throughout development and implementation.
- A framework allows us to create and collect evidence in a principled way.
- Following is an example outlined in Dogan, E., Hauger, J. & Maliszewski, C. (2014).

A validity evidence framework based on the Standards

- Framework by first dividing the assessment development and implementation period into four phases:
  - Phase I: Defining measurement targets, item and test development
  - Phase II: Test delivery and administration
  - Phase III: Scoring, scaling, standard setting
  - Phase IV: Reporting, interpretation and use of results

A validity evidence framework based on the Standards

- For each phase we identified necessary conditions and outcomes that need to be realized so that we are on track to developing an assessment system that will allow valid interpretations and uses.
- We documented **empirical** and **procedural validity evidence (**Haladyna, 2006) for each condition and outcome and provided reference to the relevant *Standard*.

A validity evidence framework based on the Standards

Phase I: Defining Measurement Targets, Item and Test Development

- **1-A:** The purposes of the assessments are clear to all stake holders.

  Relevant standards: 1.1

- **1-B:** Test specifications and design documents are clear about what knowledge and skills are able to be assessed, the scope of the domain, the definition of competence, and the claims the assessments will be used to support.

  Relevant standards: 1.2, 3.1, 3.3

- **1-C:** Items are free of bias and accessible.

  Relevant standards: 7.4, 7.7, 9.1, 9.2, 10.1

A validity evidence framework based on the Standards

Phase I: Defining Measurement Targets, Item and Test Development

- **1-D:** Items measure the intended constructs and elicit behavior that can be used as evidence in supporting the intended claims.

  Relevant standards: 1.1, 1.8, 13.3

- **1-E:** The item pool as a whole and each test form represents the blueprint and covers the entire range of student performance (e.g., low/high-achieving students).

  Relevant standards: 1.6, 3.2, 3.11, 13.3

- **1-F:** Items with high psychometric quality (e.g., high discrimination/low guessing parameters; high precision; lack of differential functioning) are identified during field testing using representative samples of examinees.

  Relevant standards: 3.3, 3.9, 7.3

A validity evidence framework based on the Standards

Sources/Evidence of Procedural Validity for Phase I

- PARCC's Application for the Race to the Top Assessment Grant (PARCC, 2010)

  Supported conditions/outcome: 1-A (description of purposes)

- PARCC Model Content Frameworks (PARCC, 2012)

  Supported conditions/outcome: 1-B (scope of domain)

- Performance-Level Descriptors (PLDs) (PARCC, 2013b)

  Supported conditions/outcome: 1-B (scope of domain)

- Summative Assessment Specifications (PARCC, 2011)

  Supported conditions/outcome: 1-B (scope of domain), 1-E (blueprint and scale coverage)

- Cognitive Complexity Framework (Ferrera, et. al., 2014)

  Supported conditions/outcome: 1-B (scope of domain), 1-E (blueprint and scale coverage)

A validity evidence framework based on the Standards

Sources/Evidence of Validity for Phase I

- **Study 1: Accessibility Studies - English Language Learners (ELLs), Students with Disabilities, and Grade 3 Students (Laitusis, et. al., 2013)**

  Supported conditions/outcome: 1-C (accessibility)

  Source of validity evidence: Construct Validity; Fairness

- **Study 2: Student Task Interaction Study (Tong & Kotloff, 2013)**

  Supported conditions/outcome: 1-D (intended constructs)

  Source of validity evidence: Response processes

- **Study 3: Quality of Reasoning and Modeling Items in Mathematics (Kotloff, King, & Cline, 2013)**

  Supported conditions/outcome: 1-D (intended constructs)

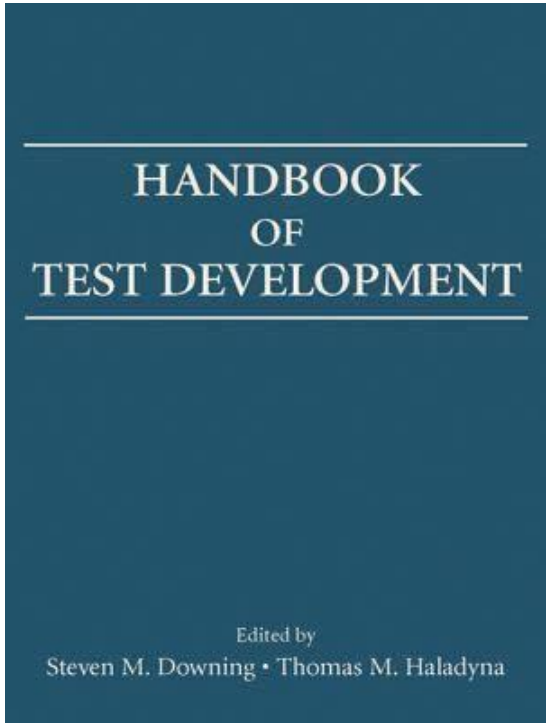
  Source of validity evidence: Test content, Response processes

- **Study 4: Use of Evidence-Based Selected Response Items in Measuring Reading Comprehension (Pearson, 2013a)**

  Supported conditions/outcome: 1-D (intended constructs)

  Source of validity evidence: Response processes

Additional resources



SCASS TILSA
Technical Issues in Large-Scale Assessment
A State Collaborative on Assessment and Student Standards

QUALITY CONTROL CHECKLIST
FOR
PROCESSING, SCORING, AND REPORTING
January 2003

SCASS TILSA
Technical Issues in Large-Scale Assessment
A State Collaborative on Assessment and Student Standards

QUALITY CONTROL CHECKLIST
FOR
ITEM DEVELOPMENT AND TEST FORM CONSTRUCTION
July 2003

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Camara, W. (2014). Standards for Educational and Psychological Testing: Historical Notes. [Public briefing]. Retrieved from http://www.aera.net/Portals/38/docs/Outreach/Standards_Hill_Briefing_Slides_FINAL.pdf?timestamp=1410876719244

Carr, P., Dogan, E., Tirre, W., & Walton, E. (2007). Large-Scale Indicator Assessments: What Every Educational Policymaker Should Know. *Yearbook of the National Society for the Study of Education*, *106*(1), 321-339.

Dogan, E., Hauger, J. & Maliszewski, C. (2014). Empirical and Procedural Validity Evidence in Development and Implementation of PARCC Assessments. In Lissitz, R. W. (Editor), *The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement*. Charlotte: Information Age Publishing Inc.

References

Haladyna, T.M. (2006). Roles and importance of validity studies. In Downing, S.M., & Haladyna, T.M. (Eds.), *Handbook of Test Development* (pp. 739-755). Mahwah, NJ: LEA.

Mazzeo, J., Lazer, S., & Zieky, M.J. (2006). Monitoring Educational Progress with Group-Score Assessments. In Brennan, R. L. (Ed.). *Educational measurement* (4th ed.). Westport, CT: ACE/Praeger Publishers.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50, 741-749.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing? Educational Measurement: Issues and Practice*, *33*(4), 4–12.

Wise, L. (2015, April). *Test Design and Development Following the Standards for Educational and Psychological Testing*. Presentation at annual meeting of the National Council on Measurement in Education. Chicago, IL.