APR Applied Personnel Research
Consulting and Expert Witness Services

Test Construction Guidance in the New (2014) Testing Standards

Joel P. Wiesen, Ph.D.
jpw@jpwphd.com

International Personnel Assessment Council
Annual Conference, July 21, 2015
Atlanta, GA

Overview of this Presentation

Some highlights from the new *Standards*

Test Specifications

Requirements for:

- Validation
- Reliability
- Documentation
- Fairness

Select topics (e.g., SMEs, grading, VG)

Sprinkling of definitions and tips

Description of *Standards*-related job aid

- Definitions, Checklist, Unanswered questions

Highlights of New *Standards*

- More extensive and voluminous than before
- What's new
- Do we have to rush to read them?

What's New and Different

- Chapter titles very similar to 1999 edition
- Some huge differences in content
  - Terminology
  - Requirements
  - Concepts
  - Emphasis
- Chapter introductions may be as important as numbered standards

Many Huge Differences

- New views of testing
- New and changed terminology
- Requirements for test plan
- Requirements for fairness
- Requirements for reliability
- Requirements for documentation

One New View of Testing

- "…if … excluding some … that could readily be assessed has a noticeable impact on selection rates … (e.g., subgroup differences are found to be smaller on excluded components …), the intended interpretation … predicting job performance in a comparable manner for all groups … would be rendered invalid." (pg 21, col 1, par 1)

Another New View

- "…consequences can influence a decision about test use, even though the consequence is independent of the validity…" (pg 21, col 1, par 2)

Fairness is Foundational

- Fairness is … "an overriding, foundational concern…"
  (pg 49, col 2, par 1)
- "Fairness is … central to the validity and comparability of the interpretation of test scores for intended uses."
  (pg 63, col 1, last par)

Change in Emphasis

Word	Frequency	% Change from 1999
content	299	87%
construct	305	82%
interpretations	302	78%
fairness	139	65%
reliability	284	61%
assessment	262	53%
scores	1,128	50%
procedures	249	49%
Total	128,475	25%

Terminology

- Propositions
  - "Decisions about what types of evidence are important for the validation argument in each instance can be clarified by developing a **set of propositions** or **claims** that support the proposed interpretation for the particular purpose of testing." [emphasis added]
    (pg 12, col 1, par 2)

Standards Sets the Bar High

- New requirements (essential, must, need to, every effort, important, should, desirable, avoid)
- Many details needed
  - I anticipate LONG validation reports.
- Some (many?) standards not clear with respect to what is expected or will be accepted

Example: a Set of Propositions

- "…certain skills are prerequisite…;
- …the content domain of the test is consistent with these prerequisite skills;
- …test scores can be generalized across relevant sets of items;
- …test scores are not unduly influenced by … variables, such as writing ability;

[continued on next screen]

Example: a Set of Propositions

- … success … can be validly assessed;
- … test takers with high scores … will be more successful … than …”
  (pg 12, col 1, par 2)

Propositions are Independent

- "…given use typically depends on more than one proposition, strong evidence in support of one part of the interpretation in no way diminishes the need for evidence to support other parts of the interpretation."

Propositions are Independent

- "… when an employment test is … considered for selection, a strong predictor-criterion relationship in an employment setting is ordinarily not sufficient to justify use of the test. … also consider the appropriateness … of the criterion … and … support for the proposed interpretation across groups."
  (pg 13, col 1, last par; pg 12 , col 1, par 2)

Propositions in Standard 1.2

- "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory…"
- "Comment: The rationale should indicate what propositions are necessary to investigate the intended interpretation." (Std 1.2, pg 23)

Example of Propositions

- Propositions for intended score interpretations (created for this IIPAC presentation):
- 1. Passing candidates will be able to do the job at a minimally acceptable level or better.
- 2. Highest ranked candidates will have more of the tested abilities than lower ranked candidates.
- 3. Highest ranked candidates will be able to do the job better than lower ranked candidates.

Stronger Propositions

- Props. for intended score interpretations:
- 1. Passing candidates will do the job at a minimally acceptable level or better.
- 2. **Higher** ranked candidates will have more of the tested abilities than lower ranked candidates.
- 3. **Higher** ranked candidates will be more successful than lower ranked candidates.
- **4.** The constructs being assessed do not vary over occasions. (pg 33, col 2, par 3)

Supporting These Propositions

1. If passing means competent, we may need to reconsider compensatory weighting of KSAs in M/C exam to be logically compelling.
2. This seems to require explicit showing that the test is a representative sample of KSA.
3. This may require a showing that the KSAs that drive performance are tested- Not necessarily the case for a Lieutenant exam if test is basically the same as for Sergeant.
4. Which reliability measure supports this claim?

Standards on Validation

A. Detailed test specifications
B. Forms of validity evidence

Terminology

- AKA:
- Test Plan
- Test Design Plan (pg 75, col 1, par 1)
- Test Outline

Detailed Test Specifications

A. Content specifications (pg 76)
B. Format specifications (pg 76)
C. Test length specifications (pg 79)
D. Psychometric specifications (pg 79)
E. Scoring specifications (pg 79)
F. Test administration specifications (pg 80)

Test Specification Overview

- "Documentation of the purpose and intended uses of a test as well as of the test's content, format, length, psychometric characteristics (of the items and test overall), delivery mode, administration, scoring, and score reporting." (Glossary, pg 225)
- Can be in multiple documents

Content Specifications

- "Describes content in detail"
- "logical or empirical analyses of the adequacy with which the test content represents the content domain"
- "relevance of the content domain to the proposed interpretation of test scores." (pg 14 col 2 par 1)

Content Specifications

- Term "content specification" not in glossary
- Content specification = content framework
  (pg 76, col 2, par 2)

Level of Detail

- "The adequacy and usefulness of test interpretations depend on the rigor with which the ... domain represented by the test ... defined and explicated."
  (Std 4.1 Comment, pg 85)

Content Specifications

- KSAPs now expanded:
  "knowledge, skills, abilities, traits, interests, processes, competencies, or characteristics"
  (pg 11, col 2, par 2)

Content Specifications

- "Specifications should be sufficient to allow experts to judge the comparability of different sets of simulation tasks included in alternate forms."
  (pg 78, col 1, par 3)

Terminology

- Construct domain
  - "The set of interrelated attributes (e.g., behaviors, attitudes, values) that are included under a construct's label." (Glossary, pg 217)

Format Specifications

- "Format specifications delineate the format of items … the response format … and the type of scoring procedures. … Format specifications should include a rationale for how the chosen format supports the validity, reliability, and fairness of intended uses of the resulting scores."
  (pg 76, col 2, par 3 to pg 77, col 1, par 1)

Terminology

- Construct
  - The concept or characteristic that a test is designed to measure (pg 11, col 1, last par)
  - "The concept or characteristic that a test is designed to measure." (Glossary, pg 217)

Test Length Specifications

- "Test developers frequently follow test blueprints that specify the number of items for each content area to be included in each test form." (pg 79, col 1, par 1)

Psychometric Specifications

- "Psychometric specifications indicate **desired statistical properties** of items (e.g., difficulty, discrimination, and inter-item correlations) … the desired statistical properties … including the nature of the reporting scale, test difficulty and precision, and the distribution of items across content or cognitive categories." (pg 79, col 1, par 2)

Scoring Specifications

- "Scoring rubrics specify the criteria for evaluating performance and may vary in the degree of judgment entailed, the number of score levels employed, and the ways in which criteria for each score level are described." (pg 79, col 2, par 1)

Scoring Specifications

- "When scoring … human judgment, the scoring specifications should describe … how scorers are to be trained and monitored, how scoring discrepancies are to be identified and resolved, and **how the absence of bias in scorer judgment is to be checked**."
  (pg 79, col 2, par 3)

Extended-Response Item Scoring

- "...for extended-response items ... Test developers **must** identify responses that illustrate **each scoring level**, for use in training and checking scorers. Developers also identify responses at the borders between adjacent score levels for use in more detailed discussions during scorer training." (pg 82, col 2, par 3)

Extended-Response Scoring

- "Providing multiple examples of responses at each score level for use in training scorers and monitoring scoring consistency is also common practice, although these are typically added to scoring specifications during item development and tryouts." (Std 4.18, Comment)

Subjective Judgments

- "…evidence should … on both interrater consistency in scoring and within-examinee consistency over repeated measurements." (Std 2.7)
- "high interrater consistency does not imply high examinee consistency from task to task. Therefore, interrater agreement does not guarantee high reliability of examinee scores."
  (Std 2.7)

Monitor Assessors, Ongoing

- "It is essential that adequate training and calibration of scorers be carried out and **monitored throughout the scoring process** to support the consistency of scorers' ratings for individuals from **relevant subgroups**."
  (3.8 Comment pg 66)

Monitor Assessors… (continued)

- "…a scoring rubric … might reserve the highest score level for test takers who provide more information or elaboration than was actually requested. … test takers who simply follow instructions … earn lower scores; thus, characteristics of the individuals become construct-irrelevant components of the test scores."

  (pg 56 col 2 par 1)

Rater Training Evaluation

- For extended-response item scoring:
- "Statistical analyses of scoring consistency and accuracy (**agreement with scores assigned by experts**) should be included in the analysis of tryout data."
  (pg 82 col 2 par 3)
- "Specifications should also describe processes for assessing scorer consistency and **potential drift over time in raters' scoring.**"
  (Std 4.20 scoring complex responses)

Index of Rater Agreement

- "The basis for determining scoring consistency (e.g., percentage of exact agreement, percentage within one score point, or some other index of agreement) should be indicated."
  (Std 4.20 Comment)
- "...should include standards for checking scorer accuracy during training and operational scoring."
  (Std 4.21 Comment)

Calibration Candidates

- How to assess drift in scoring over time?
- Perhaps use "calibration candidates"
  - Video recordings of mock candidates whose responses have been determined by the test developing SMEs
  - Have these graded as candidates 1, 20, 50, 100, etc.

Some Interviews Exempt!

- "Some assessments conducted in employment settings, such as unstructured job interviews for which no claim of predictive validity is made, are nonstandardized in nature, and it is generally not feasible to apply standards to such assessments."
  (pg 169 col 2 par 1)

Test Administration Specifications

- For CAT:
  "… to ensure that the set of items administered to each test taker meets all of the requirements of the test specifications."
  (pg 81, col 1, par 1)

Forms of Validity Evidence

(a) Content-Oriented Evidence

(b) Evidence Regarding Cognitive Processes

(c) Evidence Regarding Internal Structure

(d) Evidence Regarding Relationships With Conceptually Related Constructs

(e) Evidence Regarding Relationships With Criteria

(f) Evidence Based on Consequences of Tests

(Stds 1.11 – 1.25)

Terminology

- construct ≱ domain
- "construct: The concept or characteristic that a test is designed to measure."
- "construct domain: The set of interrelated attributes … included under a construct's label."
- "content domain: The set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test, represented in detailed test specifications…" (Glossary)

Thorough and Explicit

- "Evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest."
(Std 11.2)

(a) Content-Oriented Evidence

- "Areas of the content domain that are not included among the test items could be indicated as well"
  (Standard 1.11, Comment)

Content-Oriented Evidence

- "For example … maps the items … to the content domain, illustrating the **relevance of each item** and the adequacy with which the **set of items represents the content** domain."
  (Standard 1.11, Comment)

Content-Oriented Evidence

- "The match … to the targeted domain **in terms of cognitive complexity** and the **accessibility of the test content to all members** of the intended population are also important…"

(b) Cognitive Processes Tested

- "assumptions about the cognitive processes engaged in by test takers. … fit between the construct and the detailed nature of the performance or response actually engaged in … reasoning … instead of following a standard algorithm …"
  (pg 15, col 1, last par to 1st par, col 2)

Cognitive Processes Tested

- "... evidence that the cognitive processes being followed by those taking the assessment are consistent with the construct to be measured."
  (pg 82, col 2, par 2)

Terminology

- Cognitive Labs
- "structured interviews or think-aloud protocols with selected test takers"
  (pg 82, col 2, par 2; also see Std 3.3 Comment, pg 64)
- Can use to pretest items with various subgroups when N is small

Credentialing Different

- "Criterion-related evidence is of limited applicability because credentialing examinations are not intended to predict individual performance in a specific job but rather to provide evidence that candidates have acquired the knowledge, skills…"
(pg 175 col 2 par 3)

(c) Internal Structure

- "… degree to which the relationships among test items and test components conform to the construct …"
  (pg 16, col 1, par 3)

Factor Analysis of Item Data

- "Evidence Based on Internal Structure"
- "Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct … A theory that posited unidimensionality would call for evidence of item homogeneity."
  (pg 16, col 1, par 3; also see Standard 1.13)

Pretesting of Items Favored

- "Items … reviewed … quality, clarity, and construct-irrelevant … reviewed for sensitivity and potential offensiveness … construct-irrelevant variance for individuals … An attempt … to avoid words and topics that may offend or otherwise disturb some test takers, if less offensive material is equally useful …" (pg 82, col 1, par 1)

No Blanket Preference for IRT

- "The IRT information function is based on the results obtained on a specific occasion or in a specific context, and therefore it does not provide an indication of generalizability over occasions or contexts."
  (pg 38, col 2, par 1)

(d) Evidence: Conceptually Related Constructs

- Comment:
- "… guard against faulty interpretations arising from spurious sources of dependency among measures, including correlated errors or shared variance due to **common methods of measurement** or common elements."
  (Std 1.16, Comment)

Inference for Employment

- "The fundamental inference … from test scores in … employment settings is … prediction: The test user wishes to make an inference from test results to some future job behavior or job outcome."
  (pg 171 col 2 last par)

(e) Evidence: With Criteria

- "...information about the suitability and technical quality of the criteria should be reported."
(Standard 1.17)

Evidence: With Criteria

- Why is extreme groups approach eschewed?
- "Note that data collections employing test takers selected for their extreme scores on one or more measures (extreme groups) typically cannot provide adequate information about the association."
  (Std 1.18, Comment)
- If we believe in linear relationships for all cognitive abilities, why not support use of the extreme groups approach?

Evidence: With Criteria

- "When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided." (Standard 1.18)

Examinee Consistency/Error

- "Important sources of random error … two broad categories: … rooted within the test takers and … external to them."
  - Within: fluctuation in motivation or attention
  - External: time of day, distractions
    (pg 36 col 2 par 2)

Examinee Consistency

- "…high interrater consistency does not imply high examinee consistency from task to task. Therefore, interrater agreement does not guarantee high reliability of examinee scores."
  (Std 2.7)

(f) Evidence Based on Consequences of Tests

- Unintended consequences
  - Adverse impact on ethnic groups (pg 20, col 2, par 2)

Values for Criterion Domains

- "Decisions about test use are often influenced by additional considerations … the relative importance of selecting for one criterion domain versus others …"
  (pg 174 col 2 par 3)

Values for Criterion Domains

- "Decisions about test use are often influenced by additional considerations … concerns about applicant reactions to test content and processes…"
  (pg 174 col 2 par 3)

Values for Criterion Domains

- "Decisions about test use are often influenced by additional considerations … fairness, and policy objectives such as workforce diversity."
  (pg 174 col 2 par 3)

On Measuring Job Knowledge

- "assumptions about the cognitive processes engaged in … fit between the construct and the detailed nature of the performance or response actually engaged in … **reasoning … instead of following a standard algorithm** …" (pg 15 col 1 last par to 1st par col 2)

High Bar for Criterion Validation

- "… a strong predictor-criterion relationship in an employment setting is ordinarily not sufficient to justify use of the test. … also consider the appropriateness … of the criterion … and the consistency of the support for the proposed interpretation across groups."
  (pg 13, col 1, last par to pg 13 , col 2 par 1)

Low Bar for Validation

- "The determination that a given test interpretation for a specific purpose is warranted is based on professional judgment that the **preponderance** of the available evidence supports that interpretation." (pg 13, col 2, par 1)
- Preponderance means majority (51%)

No Hierarchy of Validity Evidence

- No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence." (Standard 1.2, Comment)

Hierarchy of Validity Evidence

- "The …inference to be drawn from test scores in most … employment settings is one of **prediction**: … inference from test … to … job outcome. … although different strategies for gathering evidence may be used, the inference … is that scores on the test can … predict … job behavior."
  (page 171, col 2, last par)

Summary of Validity Evidence

- "evidence of careful test construction;
- adequate score reliability;
- appropriate test administration and scoring;
- accurate score scaling … standard setting…;
- careful attention to fairness for all test takers"
  (pg 22, col 2, par 2)

Standards on Reliability

A. Job Analysis
B. Cut Point Determination
C. Rating essay answers, etc.
D. Accuracy within range of scores used
E. Level of reliability not specified

Careful Reliance on SMEs

- In rating candidates, are "judges … consistent with the intended interpretation of scores … ascertain whether they are, in fact, applying the appropriate criteria … studies of how observers … evaluate data … [and] the appropriateness of these processes to the intended interpretation or construct definition." (pg 15 col 2 - pg 16 col 1)

Reliability of SME Judgments

- Job analysis ratings
- "…and should report the level of agreement reached."
  (Standard 1.9 and Comment)

Reliability of SME Judgments

- Cut point ratings and rating essays
- "Systematic collection of judgments … (e.g., in setting cut scores), or in test scoring (e.g., rating of essay responses). Whenever such procedures are employed, the quality of the resulting judgments is important to the validation. Level of agreement should be specified clearly…"
  (Standard 1.9, Comment)

Reliable Cut Scores

- "clearly documented and ... defensible. ... A sufficiently large and representative group of participants should be involved to provide reasonable assurance that the expert ratings across judges are sufficiently reliable and that **the results of the judgments would not vary greatly if the process were replicated.**"
(pg 101 col 2, par 2) [emphasis added.]

Subjective Scoring

- Obtain and report evidence for reliability/precision "for each intended score use."
  (Std 2.0; also see pg 40, col 2, par 5, Stds 2.1, 2.2 , 2.6, 2.7)
- Reliability for all subgroups, as feasible
  (Std 2.11)

Reliability for Multiple Hurdles

- "For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported." (Std 2.3, pg 43)

Terminology

- Decision consistency
- Decision accuracy
- Conditional SEM at the cut score
- These are different, related concepts
  (Source: pg 40, col 1, last par)

Level of Reliability

- "However, the need for precision increases as the consequences of decisions and interpretations grow in importance. ... such as rejection or admission of a candidate..." (pg 33 col 1 par 3)
- Is "the variability associated with the error ... small compared with the observed variation in the scores (or score differences) to be estimated." (pg 34, col 2, par 4)

Documentation Standards

Standards for supporting documentation are presented in four thematic clusters:

A. Appropriate Use

B. Test Development

C. Test Administration and Scoring

D. Timeliness of Delivery of Test Documents

(pg 125, col 1, par 1-2)

All Standards "Primary"

- "... unless a standard is deemed clearly irrelevant, inappropriate, or technically infeasible for a particular use, all standards should be met, making all of them essentially 'primary' for that context."
  (pg 5, col 2, par 2)

Documentation for Small Test

- "For low-volume, unpublished tests, the documentation may be relatively brief." (Standard 7.13, Comment)

If Use a Concurrent Approach

- "The choice of a predictive or concurrent research strategy in a given domain is also usefully informed by prior research evidence regarding the extent to which predictive and concurrent studies in that domain yield the same or different results."
  (pg 17 col 2 par 2)

Validity Generalization

- "…strong basis for … [VG exists where] … the meta-analytic data base is **large**, where the meta-analytic data adequately represent the **type of situation** to which one wishes to generalize, and where correction for statistical artifacts produces a **clear and consistent pattern** of validity evidence." pg 18 col 2 par 2)

Validity Generalization

- "When a meta-analysis is used ... the test and the criterion variables in the local situation should be comparable with those in the studies summarized."
(Standard 1.22)

Low Bar for Content Validity

- For task based tests:
- "…if the test content samples job tasks with considerable fidelity … or … simulates job task content … then content-related evidence can be offered as the principal form of evidence of validity."
(Std 11.3)

Low Bar for Content Validity

- For KSA based tests:
- "…if the test content samples … specific job knowledge (e.g., information necessary to perform certain tasks) or skills required for competent performance, then content-related evidence can be offered as the principal form of evidence of validity." (Std 11.3)

More than Job Knowledge

- "… viewing a high test score as indicating overall job suitability … would be an inappropriate inference from a test measuring a single narrow, albeit relevant, domain, such as job knowledge."
  (Standard 11.4 Comment)

Standards on Fairness

Four general views of fairness

Test specifications

More Emphasis on Fairness

- Many references to fairness
- New approaches to fairness

Terminology

- Universal Design: "an approach to test design that seeks to maximize accessibility for all intended examinees."
  (Source: pg 50 col 1 par 2)

Terminology

- Accessibility: "… enable as many test takers as possible to demonstrate their standing … without being impeded by characteristics of the item that are irrelevant to the construct being measured."
  (Glossary)

Four Views of Test Fairness

A. Fair and equitable treatment of all test takers
B. Absence of measurement bias
C. Access to the constructs measured
D. Validity of individual test score interpretations for the intended use(s)

Information for Candidates

- "When test score information is released ... should describe in simple language ... the precision/reliability of the scores..." (Std 6.10)

Cautions on Validation

- "…possible distortions in meaning arising from inadequate representation of the construct and …
- to aspects of measurement, such as test format, administration conditions, or language level, that may materially limit or qualify the interpretation … for various groups of test takers." (pg 13, col 1, par 2)

Evaluating Item Fairness

- "Both qualitative and quantitative sources of evidence are important in evaluating whether items are psychometrically sound and appropriate for all relevant subgroups." (Std 3.3 Comment, pg 64)

Fairness: Access to the Constructs Measured

- Vocabulary is a potential source of interference with access
  - Artificial barrier to good test performance

B-W Differences Unintended

- "Still other consequences are unintended, and are often negative. ... As another example, a test developed to measure knowledge needed for a given job may result in lower passing rates for one group than for another. Unintended consequences merit close examination."
  (pg 19 col 2 par 1 and pg 20 col 2 par 2; also see Std 1.25)

B-W Differences: Checking

- "… unintended consequences … **especially important** to check that these consequences do not arise from construct-irrelevant components or construct underrepresentation … may also lead to reconsideration of the appropriateness of the construct in question."
(Std 1.25)

New Approaches to Fairness

- Process Studies
- Meta analysis by sub-group

Process Studies

- "Process studies involving test takers from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing test takers' test performance."
(pg 15 col 2 par 3)

High Stakes Tests

- "Adhering to the Standards becomes more critical as the stakes for the test taker and the need to protect the public increase." (pg 3, col 1, par 1)

Meta analysis by Sub Group

- "Gathering evidence about how well validity findings can be generalized across groups of test takers is an important part of the validation process."
  (pg 19 col 1 par 2)

Conclusion

Standards introduces new view of testing

Testing methodology is now better defined

Tools to help use the *Standards*

Q&A

Tools for Using the *Standards*

- Document on the web with:
- Definitions
- Checklist
- Unanswered questions
- http://aprpsych.com/standards

Q & A's

- Questions/Comments from the attendees
- Other tips you might like to suggest

Copies of Talk and Resources

- Copies of this presentation are available at http://ipacweb.org and from jpw@aprpsych.com

- Also see: http://appliedpersonnelresearch.com/papers

Reference

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.