The Standards for Educational and Psychological Testing: Zugzwang for the Practicing Professional?

Prepared for:
IPMAAC
The International Personnel Management Association Assessment Council
Newport Beach, California
June 12, 2001
Paul D. Kaiser
Kristine Smith
What the heck is Zugzwang and how can I get me some?

- It is a paradox of chess that the right to move can occasionally become an onerous obligation.
- Chess players know such situations as zugzwangs -- "zugzwang" being the German for "obligation to move."
- Simple zugzwang - one side can suffer from having the move.
- Mutual zugzwang - neither side can move without worsening its position.
Three organizations were responsible for the development of the 1999 Standards

- American Educational Research Association (AERA)
- National Council on Measurement in Education (NCME)
- American Psychological Association (APA)
Standards - 1985 and 1999 versions

The 1999 Standards

- has more background material, a greater number of standards, and an expanded glossary and index
- reflects changes in federal law and measurement trends affecting validity, etc.
- addresses professional and technical issues of test development and use
The Purpose of the Standards

◆ to promote the sound and ethical use of tests
◆ to provide assessment professionals with guidelines for the evaluation, development, and use of testing instruments
◆ to provide a frame of reference for addressing relevant issues
The Standards are not

- legislation or law
- a ‘checklist’ for evaluating the acceptability of a test or its use
Overview - Organization and Content Part One

Test Construction, Evaluation, & 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
Overview - Organization and Content Part Two

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
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
Chapter 1—Validity

The Grail--testing pointless without it

◆ refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests

◆ the most fundamental consideration in developing and evaluating tests
The Standards move away from language about types of validity to lines of validity evidence.

- test content
- response/scoring processes
- internal structure of test
- relationships to other variables
**The Standards on Validity Focus**

- on the obligations of the test developer to users, examinees, and other testing practitioners
- on user obligations to examinees
Developers owe users

Enough information to make judgments about the appropriateness of their interpretation of test scores for their intended use(s)

- population(s) for which test is appropriate
- constructs tested
- uses/interpretations NOT intended or recommended
Developers owe users

- content descriptions, domains, criticality
- qualifications of experts/judges/raters
- rating/scoring procedures
- population/situation variables involved in validation
Developers owe users

- quality of criteria
- statistical adjustments made
- relation of local situation to meta-analytic variables used
Developers owe examinees reasonable assurance that tests will not be used improperly

- use of content/constructs appropriate to recommended and/or intended use(s)
- adequate warning to users against uses NOT recommended or intended
- accuracy in criterion validation, when performed
- investigation of unintended/unexpected outcomes (e.g., DIF)
Developers owe other practitioners sound practices and evidence, accurately and adequately described

- validation samples in relevant detail
- processes/procedures in adequate detail
- criteria adequately described
- statistical evidence and descriptions of any adjustments
Users owe examinees appropriate use(s) of tests

- following test developers’ guidelines on usage of the test
- validating new uses of a test
Obligations under Standards

Developer should help User:
- use test results properly; ‘head off’ obvious improper uses
- understand test’s value, limitations
- understand degree to which test is ‘proven’

User owes Examinees:
- Proper, fair use/interpretation
- Avoiding misuse of test
Chapter 2—Reliability

- *The test* - a system for collecting examples of an individual's work or behaviors in a particular area.
- *A scoring procedure* - enables the examiner to quantify, evaluate, and interpret the behavior or work samples.
Reliability

- Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups.

- The Standards address almost exclusively the test developers’ obligations to provide information to test users.
Reliability

Consistency, replicability of measurement on repeated use with individuals or groups

- correlation coefficient
- SEM
- interrater reliability
To enable the users to make informed judgments about the test and interpretations of test scores:

- estimates of the reliability and SEM of all scores to be interpreted
- estimates of SEM in both original scale units and the units of any derived scores
- method(s) used to compute reliability
Reliability — Developers Owe Users

- statistical adjustments made to reliability estimates
- composite reliability estimates for multi-factor tests
- estimates of inter-rater reliability, when relevant
- estimate of “local” reliability of “locally” scored test
- estimate replicability of “categorical” classifications (e.g., pass/fail)
Obligations under Standards

Developer to User:
- accurate and appropriate reliability estimate
- report how estimated, including adjustments

User to Examinee:
- do not base decisions on unreliable test
Test Development

- Process of producing a measure of some aspect of an individual's KSA’s
- Guided by the stated purpose of the test
Test Development and Revision

- Primer on how to develop tests
- Definitions of terms and concepts involved in test development (e.g., norm/criterion referencing, holistic scoring)
Test Development and Revision

Standards focus on

- Stating the purpose
- Developing a framework
- Developing test specifications
- Developing and evaluating items
- Assembling the test
Test Development and Revision

- Document what you do
- Describe the purpose of the test and the domains to be covered (i.e., constructs, SKAPs)
- Document the test specifications
- Have SMEs review the test specifications and test items
Test Development and Revision

- Document procedures for interpreting scores
- Document procedures used to write, review, pretest, and select items
- Where scores are derived from differential weighting of test items, document the rationale and process used
Test Development and Revision

- Document rating scales for constructed response (i.e., short answer, essay, and performance) tests.

- Where tests have time limits, examine the extent to which speed is a factor in test performance.

- Indicate clearly to test takers when a test is for research purposes only.
Obligations under Standards

Document everything involved

- development of test specifications
- SME review of specifications and items
- procedures for interpreting scores
- procedures to write, pretest, select items
- rationale and process for any weightings
- rating scales for constructed responses
- alternate form specifications meet originals
Focus of Standards is on converting raw scores to some form of scaled score in order to enhance the scores’ interpretability and meaning.

Scaling a test means choosing a scoring formula or set of formulas to accomplish the conversion.
Scales, Norms, and Score Comparability

- Converting raw to scaled scores
- Terms and concepts in scaling, equating, etc.
- Bases for passpoints and bands
- Alternate forms
Some Definitions

- Raw score
- Scaled or derived score
- Standards or cut scores
Basis for cut scores

- number to be hired or promoted
- empirical research
- SME judgment (e.g., Angoff, Nedelsky)
- Norm-referenced or Criterion-referenced
Scales, Norms, and Score Comparability

- Alternate forms
- Equating
- Adaptive tests
- Linkage, calibration, concordance, projection, moderation, and anchoring
Clearly explain scales used to convert scores

If specific misinterpretations of score scales are likely, forewarn the test users

Describe norming population and samples clearly

When norms are used to characterize groups (in contrast to individuals), the statistics used to describe the group need to be clearly explained.
Clearly explain the rationale for the criterion-referenced score interpretation.

Explain and provide evidence to support the equivalence of scores on alternate forms.
Obligations under Standards

- Explain scales and meanings of scaled scores
- Explain interpretations of any scores
- Describe norming populations, process
- Explain rationale for criterion-referenced interpretation
- Document rationale for cut scores/bands
- Let SMEs work as SMEs
Chapter 5–Test Administration, Scoring, and Reporting

- Standardized instructions/procedures increase reliability, score interpretability
- Standardization and test security helps insure test fairness
- Disability may require modified administration
- Examinees should be given enough information to interpret their scores
**Obligations under Standards**

- Test administration should be standardized
- Modifications or disruptions of administration procedures or scoring should be documented
- Test takers should be informed of procedures for requesting and receiving accommodations in advance of testing
- The testing environment should furnish reasonable comfort with minimal distractions
Obligations under Standards

- Instructions should indicate how to make responses and use unfamiliar equipment
- Eliminate opportunities for test takers to attain scores by fraudulent means
- Test users are responsible for protecting the security of test materials at all times
Obligations under Standards

- Protect the confidential nature of the reported scores
- Material errors found in test scores should be corrected ASAP
- Protect test security; prevent cheating, fraud
- Score accurately
Chapter 6—Supporting Documentation for Tests

◆ Objective is informed decisions by user
◆ Primary communication channel to users
◆ Should be complete, accurate, current, clear
◆ Specify nature of test, use, development process, scoring, interpretation, validity, reliability, scaling, norming, administration
Obligations under Standards

- Develop, as needed, appropriate user guides
- Document all studies and analytical procedures
- Keep documentation on all job analysis activities and test development steps
- Document the item selection procedures (Job expert use, pre-testing, etc.) used for the test.
Obligations under Standards

◆ Keep files of job specifications, announcements (minimum qualifications, duties statements, etc.) for the test.

◆ Keep documentation on all statistical analyses and passpoint setting procedures performed on the test.

◆ Keep files of any specific studies done on the test.
Obligations under Standards

- Keep files of ethnic and gender item analysis, etc.

- If the separate answer sheet response method is used and computerized tests are also developed, determine if scores are interchangeable or if the response method used affects scores.

- Test booklets and all test related analysis materials should be properly and accurately dated.
Obligations under Standards

- documentation understandable by user
- document uses and warn on misuses
- give reliability, validity data, if any
- specify administrator qualifications
- prove alternate forms really are
- support interpretations
Chapter 7–Fairness in Testing and Test Use

Focus of Standards is on

- responsibilities of those who make, use, and interpret tests
- those aspects that are characterized by some measure of professional and technical consensus
Fairness in Testing and Test Use

4 characterizations of fairness

1. no bias
2. equitable treatment in process
3. equal outcomes
4. equal opportunity to learn content
bias: construct irrelevancies which lower or raise scores for identifiable groups

absolute fairness to all impossible

(3) almost entirely repudiated

Differential Item Functioning (DIF)
Sensitivity Review Panels

- Pre-Test
- Post-Test (DIF)
Obligations under Standards

- if scores differ, get validity evidence for each subgroup
- only use test for group if valid for group
- conduct sensitivity reviews
Obligations under Standards

- conduct DIF studies when feasible
- keep verbal level to minimum valid level
- check that group differences are not based on content irrelevancies or construct under-representation
- equitable treatment during testing
Chapter 8–The Rights and Responsibilities of Test Takers

- Fairness issues unique to the interests of the individual test taker
- Test takers have responsibilities
The Rights and Responsibilities of Test Takers

- Fair treatment promotes validity
- Test takers should get info on: nature of test, use, confidentiality, available accommodations
- Test takers have responsibilities to: prepare for test, follow directions, answer honestly, not cheat, not steal material, not violate test security
The Rights and Responsibilities of Test Takers

- Information about the test that is available to any test taker should be available to all test takers
- Test takers should be informed about test content, including subject area, topics covered, and item formats
- Scores of individuals should be kept confidential
The Rights and Responsibilities of Test Takers

- Data files should be adequately protected from improper disclosure
- Test takers should be made aware that any form of cheating is inappropriate and that such behavior may result in sanctions
- Any form of cheating or behavior should be investigated promptly
Obligations under Standards

- Make same information on test available to all
- Inform test takers of content and test format
- Maintain confidentiality
- Warn of consequences of cheating
- Investigate possible cheating, fast, fairly, with appeal available for disqualification
Centers on translated tests

Any test using language is partly a test of language skill

Lack of it may invalidate measure of KSA

OK to test in language and at level needed for job

Similar issues involved in testing some disabled candidates
Testing Individuals of Diverse Linguistic Backgrounds

- Design tests to reduce threats to the reliability and validity of test score inferences that may arise from language differences.
- Generally, the test should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment.
Bilingual individuals can vary considerably in their ability to speak, write, and read in each language.

These abilities are affected by the social or functional situations of communication.
Testing Individuals of Diverse Linguistic Backgrounds

- Language level needed for the test should not exceed the level needed to meet work requirements.
- Issues associated with bilingual testing are also relevant to testing individuals who have unique linguistic characteristics due to disabilities such as deafness and/or blindness.
Obligations under Standards

- Design test to reduce invalidity based on language differences
- Give test in taker’s best language, unless language is part of assessment
- If modified test scores comparable to original, do not “flag” modified score
- Test language level not to exceed job need
Chapter 10—Testing Individuals with Disabilities

- Modification to test format, response format, timing, setting, content
- Modification to eliminate construct-irrelevant differences in performance
- Modification should not change construct
- Modification should not put those with modified test at undue advantage over “regular” test takers
Testing Individuals with Disabilities

Definitions

- Individuals with Disabilities
- Accommodation
Testing Individuals with Disabilities

Modification not appropriate under variety of circumstances

- If test designed to assess essential skills, and would fundamentally alter construct being measured
- Disability such that would not influence performance on test
- Requested modification exceeds “reasonable accommodation” for the disability
Alter the medium to present test instructions

- For visual impairments (e.g.—Braille, large print, computer administered oversize computer screens, larger fonts)
- For hearing disability (e.g.—sign communication or writing)
Testing Individuals with Disabilities

Modifying Response Format

- Allow use of preferred communication modality
  - Severe language deficit – can point to response
  - Manual disability – amanuensis, tape recorder, computer keyboard, Braillewriter
Testing Individuals with Disabilities

Modifying Timing

- Breaks during testing
- Extended time
- Extended testing over several days
Testing Individuals with Disabilities

Modifying Test Setting

- Individualized testing
- Location wheelchair accessible
- Tables and chairs
- Altered lighting conditions
Testing Individuals with Disabilities

Using only Portions of Test

- Waive oral test for hearing disabled
- Substitute Tests or Alternate Assessments
Obligations under Standards

- Take steps to ensure score differences based on construct, not disability
- Have knowledge/expertise on test/disability interaction
- Pilot test
- Document modifications, effects
Obligations under Standards

- Set empirical time limits, etc.
- Validate on test takers with disability
- Use an appropriate modification
- Alert users to relevant changes only
Chapter 11—The Responsibilities of Test Users

- Primary focus is to protect those tested from improper use of tests
- Aimed at users who select, give, apply tests
- Covers issues for users to consider when performing those activities
Obligations under Standards

- Use appropriate test
- Allow only trained persons to pick, give, interpret tests
- Know how test adds value to decisions
- Give timely, understandable results to examinees
Chapter 12—Psychological Testing and Assessment

Four uses of psychological testing

- diagnosis
- intervention planning & evaluation
- legal & government decisions
- personal awareness, etc.
Psychological testing is used in employment testing to:

- answer specific questions about a client’s psychological functioning during a particular time interval
- predict a client’s psychological functioning in the future
Obligations under Standards

- Users--stay in areas of competence
- “Know your tests,” pick right one(s)
- If combination of tests, use ones that “work” together
Obligations under Standards

- For differential diagnosis, test must differentiate
- Provide pretest information and results
- Train administrators and scorers
- Maintain security and confidentiality
Chapter 13—Educational Testing and Assessment

- Testing in formal educational settings
- 3 areas: routine, system-wide; selection for higher ed.; individualized/special
- Educational tests: plot KSAs vs. goals
- Stakes: effects on test-takers; the higher the stakes, the more evidence of quality needed
Obligations under Standards

- Specify uses of mandated tests
- Show quality of multiuse tests for uses
- Norm locally
- Give students opportunity to learn, and to be retested if stakes high
Obligations under Standards

- Validate placement or promotion tests
- Have qualified monitors, supervisors, score interpreters
- Test preparation should not affect validity
- Score reports: test date, SEM, interpretation.
Chapter 14—Testing in Employment and Credentialing

- In employment: selection, placement, promotion
- Context: candidate pool, screening in or out, sole determiner or not, applicant count, selection ratio
- Credentialing: standards for practitioner
- Test should be valid, cut score appropriate
Professional or Occupational Credentialing

- Tests are intended to identifying practitioners who have met particular standards.
- Qualifications typically include educational requirements, supervised experience, and attainment of a passing score on tests.
Test design requires a definition of the occupation so that persons can be clearly identified as engaging in the activity.

Validation strategies rely primarily on content-related evidence.

Verifying the appropriateness of the cut score is the critical element.
When designing and evaluating an employment test, contextual features such as the following should be considered:

- internal vs. external candidate pool
- untrained vs. specialized jobs
Testing in Employment and Credentialing

- short-term vs. long-term focus
- screen in vs. screen out
- mechanical vs. judgmental decision making
- size of applicant pool relative to the number of job openings
Obligations under Standards

- Validation: congruent with testing objectives
- Important work behaviors as criteria
- Base content evidence on thorough, explicit definition of domain, from JA
- Credentials: set cut score for performance needed, not numbers
Program Evaluation: process to judge need for, value of program

Typically infers from tests designed for other uses, so is secondary data analysis

Tests so used should meet Standards
Obligations under Standards

- Show quality of multiuse tests for each
- Define and validate any change scores
- Monitor impact, minimize negatives of mandated tests; maintain test integrity
- Inform legitimately interested: admin, scoring, score retention, release conditions
- Prevent misinterpretation of scores