AERA Mini-course: Using the *AERA/APA/NCME Standards for Educational and Psychological Testing* to Improve the Quality of Educational Research

Chapter 3, Fairness in Testing

AERA 2017 Annual Meeting

Linda Cook

Organization of Presentation

- Background for Fairness in Testing Chapter
- Fairness in Testing Standards
- Review of Scenarios

Background

- Our Vision for the Fairness Chapter
  - Fairness is an Integral Part of the Validity of Score Interpretations
  - Fairness is a Fundamental Principle Underlying all Steps in the Testing Process
    - Test design, development, administration, scoring, test use, test score interpretation
  - Fairness is Not Something You Think About "After the Fact"

Background

- Our Vision (continued)
  - Principles of Fairness Apply Regardless of the Individual or Subgroup Characteristics
  - Fairness is a Fundamental Right of all Individuals and Subgroups in the Intended Test Population

Background

- Background Section:
  - General Views of Fairness
  - Threats to Fair and Valid Interpretations of Test Scores
  - Minimizing Construct –Irrelevant Components Through Test Design and Testing Adaptations
- Standards
  - 20 Standards Organized into Four Clusters

Four Clusters for the Fairness Standards

1. Test Design, Development, Administration and Scoring Procedures that Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups
2. Validity of Test Score Interpretations for Intended Uses for the Intended Examinee Populations
3. Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretations of Scores for Their Intended Uses
4. Safeguards Against Inappropriate Score Interpretations for Intended Uses

Overarching Fairness Standard

- All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

Fairness in Testing Standards

- Cluster 1: Test Design, Development, Administration, and Scoring Procedures that Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups.
  - Contains 5 Standards
  - Standards stress importance of designing and developing tests free of construct-irrelevant barriers for widest range of test takers
  - Standards call for test developers to evaluate test appropriateness for subgroups during development and piloting
  - Test developers should document provisions made in test design, development, administration and scoring to support fairness
  - Test security is important to fairness in testing

Standard 3.1 Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.

Comments

- Design all steps in testing process to promote valid interpretations for widest range of individuals and subgroups in intended population
- Consider creating test using principles of Universal Design
    - Take into account characteristics of all intended test takers
    - Define constructs precisely
    - Avoid formats or characteristics of items and tests that may compromise valid score interpretations for individuals or subgroups
    - Provide simple, clear and intuitive testing procedures and instructions
- The ultimate goal of test design is to design a process that will, to the extent possible, remove potential barriers to the measurement of the intended construct for all individuals in the intended test population

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 characteristics, such as linguistic, communicative, cognitive, cultural, physical or other characteristics.

Comments

- Language used in tests should be consistent with the purpose of the test and familiar to as wide a range of test takers as possible
- Test developers should avoid language that has different meaning for subgroups
- Level of language proficiency should be minimum required to meet work or credentialing requirements or to represent the target construct
- In work situations, the modality used to assess language proficiency (oral, written, spoken) should be consistent with job requirements

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.

Comments

- Test developers should use individuals from relevant subgroups in pilot and field test samples
  - Analyses should focus on detecting aspects of test design, content and format that might distort interpretations of scores for subgroups and individuals
  - If sample sizes permit, desirable to carry out analyses separately by subgroup
  - If sample sizes not large enough, could accumulate data over time or use small sample techniques
  - Sensitivity reviews can be effective ways to guard against construct-irrelevant language and images being used in the test

Standard 3.4 Test takers should receive comparable treatment during the test administration and scoring process.

Comments

- Important to use standardized protocols for test administration, scoring, and test security
  - Technology-based testing adds extra concerns for standardization in administration and scoring
  - If some test takers use computers that are slower or have poor screen resolution they could be unfairly disadvantaged
- Good test security procedures are an essential part of a fair test administration

Standard 3.5 Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.

Comments

- Test developers should document how they minimized construct-irrelevant barriers in testing process
- Studies carried to examine the reliability/precision of scores and validity of score interpretations should be documented
- Special scoring, administration or reporting procedures should be documented

Fairness in Testing Standards

- Cluster 2: Validity of Test Score Interpretations for Intended Uses for the Intended examinee Population.
  - Contains 3 Standards
  - Standards stress responsibility test developers and score users have for examining validity of score interpretations for subgroups when credible evidence or theory suggests scores may be biased
  - Standards also stress responsibility of test developers and score users for investigating the possibility of differential prediction for subgroups
  - Test developers and score users are held responsible for investigating the validity of the scoring process for constructed response items

Standard 3.6 Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.

Comments

- Simple differences in scores between subgroups do not necessarily indicate lack of fairness but differences indicate need for follow up
- Reasons for subgroup differences can be investigated using qualitative and/or quantitative methods
- Try to accumulate data over test administrations if sample sizes are too small for standard psychometric analyses
- Qualitative studies such as focus groups, expert reviews and cognitive labs can be quite useful

Standard 3.7 When criterion-related validity evidence is used as a basis for test score-based predictions of future performance and sample sizes are sufficient, test developers and/or users are responsible for evaluating the possibility of differential prediction for relevant subgroups for which there is prior evidence or theory suggesting differential prediction.

Comments

•Differential prediction studies should be carried out for subgroups when tests are used to predict future performance

Standard 3.8 When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.

Comments

.

- Expectations and perceptions of human scorers can introduce construct-irrelevant variance in scores from constructed response tests
- Procedures for human scoring should be designed so that the scores are not influenced by the perceptions and predispositions of the scorers
- Human scorers should be trained, calibrated and monitored to support consistency of ratings for individuals from different subgroups
- When sample sizes are large enough the precision and accuracy of scores for relevant subgroups should be calculated
- The validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups

Fairness in Testing Standards

- Cluster 3: Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretations of Scores for Their Intended Uses.
  - Contains 6 Standards
  - Although standards emphasize need to consider widest possible range of test takers, some test takers will still need adaptations to test or testing procedures
  - Two types of adaptations: accommodations do not change construct measured by the test; modifications do change construct measured by the test
  - Test developers and score users are responsible for developing and providing accommodations and/or modifications and for developing standard administration procedures
  - Some standards in Cluster 3 provide guidance for language accommodations/modifications

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.

Comments

- An appropriate accommodation responds to individual's characteristics but does not change construct measured by the assessment
- Test developers should document basis for concluding accommodation does not change construct assessment is measuring
- Accommodations should address individual test taker's needs
- Modifications that change the construct an assessment is measuring may be needed so that examinee can demonstrate their standing on some part of the construct measured by the assessment
- If a test is modified, the modified assessment should be treated like a newly developed assessment that needs to adhere to the standards for validity, reliability/precision, fairness, and so forth

.

Standard 3.10 When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

Comments

- Accommodations should only be used when the test taker has a documented need for the accommodation
- Test developers and/or test users should provide information about the availability of accommodations and procedures for requesting accommodations to test takers
- Instructions for administering accommodations should be clearly documented and test administrators should be trained to follow the instructions
- Administration procedures should include recording which accommodations were used for specific individuals and where relevant, for recording any deviation from standardized procedures for administering accommodations

.

Standard 3.11 When a test is changed to remove barriers to the accessibility of the construct being measured, test developers and/or users are responsible for obtaining and documenting evidence of the validity of score interpretations for intended uses of the changed test, when sample sizes permit.

Comments

- Pilot or field test accommodations with individuals from subgroups who will benefit from the accommodation
- A goal of the field test should be to investigate the comparability of inferences made from the accommodated scores and scores on the original test
- Evidence should be provided for any recommended changes to the test or testing procedures
- When tests are linguistically simplified, the test developer is responsible for providing evidence that inferences based on scores from the linguistically simplified test are comparable to inferences based on scores from the original test

.

Standard 3.12 When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for intended use.

Comments

- When multiple language versions are intended to provide comparable scores, test developers should describe in detail the method used for the test translation and adaptation and should report evidence of the validity of the scores
- Validity evidence could be in the form of empirical studies or professional judgment
- When sample sizes permit, evidence of score accuracy and precision should be provided for each subgroup that the test is intended for and the properties of the test for each group should be included in a test manual

.

Standard 3.13 A test should be administered in a language that is most relevant and appropriate to the test purpose.

Comments

•Test users should take into account the linguistic and cultural characteristics and language proficiencies of test takers who are bilingual or use multiple languages
•Identifying the most appropriate language for testing requires consideration of the context and purpose for testing
•In some cases it may be more appropriate to administer the test in the language of instruction even though the test taker may be less proficient in that language than in another language
•Determination of a test taker's most proficient language for test administration does not automatically guarantee validity of score inferences for the intended use
.

Standard 3.14 When testing requires the use of an interpreter, the interpreter should follow standardized procedures and, to the extent feasible, be sufficiently fluent in the language and content of the test and the examinee's native language and culture to translate the test and related testing materials and to explain the examinee's test responses, as necessary.

Comments

•.Examinees with limited language proficiency should ideally be tested by professionally trained bilingual/bicultural examiners

•If an interpreter is required, the test user is responsible for selecting an interpreter with reasonable qualifications

•The interpreter needs to understand importance of following standardized procedures and accurately conveying the test takers responses

•The interpreter needs to be familiar with the meaning and associated vocabularies of any technical terms that are used in the test in both languages

•Unless a test has been standardized and normed with the use of interpreters, their use needs to be viewed as an alteration that could change the measurement of the intended construct

Fairness in Testing Standards

- Cluster 4: Safeguards Against Inappropriate Score Interpretations for Intended Uses
  - Contains 6 Standards
  - The standards in Cluster 4 warn test developers, test publishers, and test users not to use tests when there is no evidence to support the use with a particular group or for a particular purpose
  - The standards caution against reporting of disaggregated test results for subgroups unless there is evidence of comparable meaning across groups
  - Standards caution against use of scores alone for diagnostic testing, high stakes outcomes, or special program placement
  - Scores should not be used for high stakes decisions if students have not had the opportunity to learn the test material

Standard 3.15 Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups.

Comments

- Test developers should include explicit statements about the applicability of the test for specific subgroups in test manuals and instructions for test interpretation
- Test developers need to present evidence of the applicability of the test for relevant subgroups and make explicit statements about possible misuses of test scores for these subgroups

Standard 3.16 When credible research indicates that test scores for some relevant subgroups are differentially affected by construct-irrelevant characteristics of the test or of the examinees, when legally permissible, test users should use the test only for those subgroups for which there is sufficient evidence of validity to support score interpretations for the intended uses.

Comments

- Tests do not always measure the same thing for different subgroups
- The decision to use a test with a particular subgroup requires a careful analysis of the validity evidence for that subgroup
- In cases where there is differential validity, test developers should provide guidance to score users about score interpretations that can be made for individuals from subgroups
- Sometimes the law may limit the extent to which a test user may evaluate groups who have taken different tests

Standard 3.17 When aggregate scores are publicly reported for relevant subgroups—for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults—test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups.

Comments

- Reporting scores for relevant subgroups is justified only if the scores have comparable meaning across the groups
- The comments caution against using terms to describe subgroups that are not sufficiently precise

Standard 3.18 In testing individuals for diagnostic and/or special program placement purposes, test users should not use test scores as the sole indicators to characterize an individual's functioning, competence, attitudes, and/or predispositions. Instead multiple sources of information should be used, alternative explanations for test performance should be considered and the professional judgment of someone familiar with the test should be brought to bear on the decision.

Comments

- Many test manuals point out that additional information should be considered in interpreting test scores
- Examples of this type of information are clinical history, medications, high school record, motivation, vocational status, age, gender, etc.
- Opportunity to learn is an important factor that should be taken into consideration
- Test users are responsible for interpreting individual scores in light of alternative explanations and/or relevant individual variables noted in the test manual

Standard 3.19 In settings where the same authority is responsible for both provision of curriculum and high-stakes decisions based on testing of examinees' curriculum mastery, examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content.

Comments

- In educational settings, students' opportunity to learn the content and skills assessed by an achievement test can seriously affect their test performance and the validity of test score interpretations for intended use for high stakes individual decisions
- If there is not a good match between curriculum and instruction and the tested constructs, students cannot be expected to do well on the test and can be unfairly disadvantaged by high stakes decisions based on the test scores

Standard 3.20 When a construct can be measured in different ways that are equal in their degree of construct representation and validity (including freedom from construct irrelevant variance), test users should consider, among other factors, evidence of subgroup differences in mean scores or in percentages of examinees whose scores exceed the cut scores, in deciding which test and/or cut scores to use.

Comments

- The comments for this standard point out that evidence of differential subgroup performance is an important factor influencing the choice between tests
- Other factors such as cost, testing time, test security, and logistical issues enter into professional judgments about test selection and use
- If the scores from two tests lead to equally valid interpretations and impose similar costs and other burdens, legal considerations may require selecting the test that minimizes subgroup differences

Scenarios: Background

- Testing organization contracted with large mid-western state to develop *State A Survey of English and Mathematics Skills* for grades 3-8
- Assessment must be challenging, innovative, aligned with content standards, suitable for diverse population of test takers and computer based
- Grade 8 mathematics test is non adaptive 2 hour test administered on computer
- Test is administered in English and contains CR and MC questions. Some questions require calculator, some require graphic manipulation. One question requires test taker to write short essay

Scenario 1

- Ying has recently moved to city in large mid-western state. She has always demonstrated strong math skills
- Her parents surprised when she scored in 85th percentile
- What might be some possible causes for Ying's apparent low performance on the test?
- Is there reason to suspect that Ying's score may not fairly represent her mathematical skills or abilities?
- What questions should be asked to evaluate whether or not the test or testing conditions are a factor in Ying's low score?
- What standard/standards might apply in this situation?

Scenario 2

- Dan is carrying out research study using scores from State A Grade 8 Mathematics Assessments
- Data file indicates some students took test with accommodations and some with modifications
- How should Dan interpret the scores for students who took test with accommodations?
- How should he interpret the scores for those who took the test with modifications?
- What evidence should Dan expect the test publisher to provide to help him interpret the scores of the modified or accommodated test?
- What should Dan do if the test publisher cannot provide him with evidence of the validity of score interpretations for the modified or accommodated test?

Scenario 3

- Average scores for African American students were lower on Grade 8 Mathematics Assessment than for other groups who took the test
- State asked a researcher from the State University to examine results of test. Mary informed state that results could be due to a number of factors
- What are some of the test features that Mary could examine to assure the state that the new test is fair to all subgroups who take it?