Evaluating Testing with
Reference to Individuals with Disabilities

Kurt F. Geisinger Ph.D.

Shue Yan University

October 2018

A Basic Introduction to the Flow

- Explanation of Goals
- This presentation focuses on how to evaluate the testing of individuals with disabilities
- First we must understand how to evaluate all testing
- Description of Test Materials
  - The kinds of information needed
  - Test development information
  - Test evaluation information
  - Test use information
- How to evaluate a test based on data
- Then consider the testing of individuals with disabilities
- How we need to evaluate their testing and assessment

Brief History of Buros Center for Testing

- Founded by Professor Oscar K. Buros in the late 1930s
- Envisioned as the *Consumer Reports* for the testing industry
- Relocated at the University of Nebraska in Lincoln (UNL) in 1979 (shortly after Professor Buros' death)
- Primary publications:
  - 20 *Mental Measurements Yearbook*s (12 at UNL)
  - 9 *Tests In Print*
  - 2 *Pruebas Publicadas en Español* (by October, 2018)

So How Do We Evaluate Tests?
Two Sources: Buros and the Standards

- The Buros Outline in the *Mental Measurements Yearbooks*
  - Description
  - Development—secondary focus of evaluation
  - Technical—primary focus of evaluation
  - Commentary
  - Conclusion

Let's Consider that Outline

- **Description**—a brief overview of the test, what it looks like, how it is used
- **Development**—Who developed the measure and how was the test developed?
- **Technical**—What research is there to justify its use, its adequacy, and its technical quality?
- **Commentary**—Overall what does the reviewer think about it? What is their evaluation?
- **Conclusion**—A brief summarization of the above

What is Missing from the Outline?

- Administration ease of the test
  - Covered in the Description section
- How test takers react to the test
- Uses to which the test should NOT be put
- Cost considerations
- Availability in other languages, for those with disabilities

Description of Tests

- Attempt to describe the test briefly:
  - What is the test?
  - To whom is it to be given?
  - For what uses is it intended?
  - How is it administered?
  - What are possible misinterpretations or ways that it should not be used?
  - What else should a person know about it?

Key Issues in the Evaluation of Tests and Assessments re: Description #1

- Depends upon the intended audience
- Access
  - Who can receive materials that describe the measure?
  - Who can receive an example copy of the test, if not anyone?
- Are there test items available for inspection?
- Are there multiple forms of the test?
- To what extent do these materials seem like a sales pitch?
- To whom is the description accessible?

Key Issues in the Evaluation of Tests and Assessments re: Description #2

- To what extent is the test available for special populations?
  - Those from different cultures
  - Those who are language minorities
  - Those with disabilities (of what types)
  - What accommodations, if any, are possible for persons with disabilities or from language minorities

Key Issues in the Evaluation of Tests and Assessments re: Description #3

- Are the proposed uses of the test explained carefully?

- Are proposed interpretations of test results (i.e., scores) also detailed?

- Are potential misuses also described?

Major Characteristics for Evaluating Tests

- **Validity,** validity, validity
  - What is validity? Test does what it is supposed to do
  - Protects the public from those lacking such knowledge!!!
- **Reliability** (reproducibility)
- **Fairness** (to protected groups, by law)

What are some of the technical issues?

- How were items composing the test evaluated?
- What reliability information is provided?
- What validity information is provided?
- Is the validity information related to the test uses per se?
- Who conducted the above research? How independent is it from the authors/developers?

What are some of the technical issues?

- How was the passing score determined? Does it differ by locale? What percentage of those taking it pass?
- Equating of forms, if any? How was it done?
- Is the test developed through classical psychometrics or item response theory? What data are available regarding the test?
- What kinds of specialized knowledge, skills or certification are needed to use the test?

Item Analysis and Selection

- What kind of pre-testing was performed?
- On what sample (size, appropriateness, representativeness to population)?
- What item analysis procedures were performed? Were they based on a pre-test or the actual examination?
- How were the relevant content domains represented?
- Who interpreted these data and made decisions?

Reliability #1

The issue of reliability is a key aspect of test evaluations

- What kind of reliability analysis was done?
- Reliability represents a family of interpretations and procedures
  - Test-retest
  - Alternate forms
  - Internal consistency (most common)

Reliability Analyses #2: SEMs

- Standard errors of measurement indicate how much an individual's score is likely to change if they take the test again (a different form?)
- What kind of information is provided to users of the tests in terms of standard errors of measurement?
- To what extent are scores likely to move across the passing score?
- Were the standard errors of measurement calculated on appropriate indices of reliability?

Reliability #3: Scales

- Are there separate scales? (Often used for formative feedback)
- If so, were reliability analyses performed for each scale?
- Was there also an overall reliability?
- Is it sensible to have separate scales and an overall score?
- Were the scales weighted in any way and if so is that taken into account in the overall analysis?
- Were the correlations among scales provided? Again, on what sample . . .

Reliability Analyses #4: Tests with Cut (Passing) Scores

- Is a reliability of decisions also provided?
- Is there an appropriate standard error of measurement provided around the cut score? Around each cut score if there are a number of them?
- Are percentages of changed decisions estimated (e.g., upon re-testing)?

Validity and Validation

- **Validity** is always based on evidence
- Tests are not valid or invalid, they fall in between
- They are potentially valid for a particular use with a particular population at a particular time
- We collect evidence on appropriate uses of tests: validation
- **Validation** is never-ending and on-going
- It is the assembly of this validation information that is the primary nature of test evaluation
- Validation in many ways has replaced validity

Validity Issues

- Historical conceptions of validity (through 1985)
  - Criterion-related validity (predictive & concurrent)
  - Content-related validity
  - Construct-related validity

- New conceptions: all validity is one
- All validity is based on evidence
- Singular validity looks a lot like construct validity

Validity Issues

- Think of validity as arguments—in a legal sense, arguments that justify uses
- Think also about possible alternate uses to which a test might be put—would it be appropriate to evaluate schools based upon how their graduates do on the measure?
- Like Reliability, it is not a resident property of the test

Criterion-related Validity: Points that Must be Made

- Often used for admissions measures
- Is there a good criterion of ultimate performance?
- How representative is this criterion? (psychiatrist example)
- Were the people tested in the study like those where the test will be used in practice?
- Who would evaluate the quality of the work?

Criterion-related Validity: Points that Must be Made

- Is the setting similar to what will be used in practice?
- Not typical for licensure examinations, but may become more common
- If various demographic groups are all to be tested
  - Were all included in the studies?
  - Were the results of these predictions comparable?

Construct Validation

- Empirically answers questions as to whether the test is measuring what it is supposed to measure
- How does it relate to other accepted measure: probably not able to be done with this measure
- Relates to how able the measure is to differentiate those with knowledge and skill from those who do no have them
- Typically the best kind of validity, but difficult for licensure measures

Construct Validation Information

- Correlations with other tests
- Group differences
- Training and experience differences
- Very difficult with a licensing test unless there is a highly valued existing measure

Content Validation

- Many validity theorists insist that validity relates to score interpretations
- That is, validation supports the interpretations we make of test scores
- So can what is on the test be validity?
- Clearly, it can relate to validity, but can it justify an interpretation of validity?

Content Validation

- Does the test cover the content that it is supposed to cover?
- The way most educational and certification tests are built
- Is the content relevant to the job at the entry level?
- Is the content representative of the knowledge and skills needed?

Case #1

- John has been diagnosed by the school psychologist in elementary school (as early as first grade) as having ***dyslexia***. Although he performs satisfactorily in Science and Mathematics, he has had great difficulty in English Language Arts throughout his education. He has always been in regular classes rather than special education. On those in-class tests that he has taken in grades 1-3, the teachers have provided him with extra time. In a few weeks, he needs to take the statewide assessment (in Reading and Mathematics) in his state. Because the state has a strong impetus to encourage all students who are able to take the standard examination, the principal, on the form that she must complete for the State Department of Education, has indicated that he can take the test, but needs additional time, and notes that this is consistent with the IEP under which he has been working. The elementary school is going to provide a special testing for those students like John with special needs. John's parents react very negatively to his exclusion from the normal classroom and insist that he take it with all the other students.

- ***As director of testing for the district, what do you do?***

Cronbach's Famous Paradigm

- Develop a test plan
- Have two teams INDEPENDENTLY build a test to those specifications
- Administer both tests in counter-balanced to an appropriate sample
- Correlate results
- Violà—Coefficient of Content Validity
- However, no one ever does this ($$$)

The Main Argument for Content Validation

- Being a competent dentist is based in large measure on knowledge
- Can inspecting the items on a test provide relevant information for this determination?
- Does the test provide for both content representativeness and content relevance?
- How are skills measured as opposed to knowledge?

A Second Argument for Content Validation

- The 2014 US Joint Technical Test Standards list content as one of five characteristics that can be looked at in validation.
- The 5 characteristics are:
  - Evidence based on **Test Content**
  - Evidence based on Response Processes
  - Evidence based on Internal Structure
  - Evidence based on Relations with Other Variables
  - Evidence based on the Consequences of Testing

Problematic Example #1

- I was a consultant to the *Psychologist Licensing Examination* in the late 70s and early 80s
- Reasonably small portion of the total test was devoted to Research Design, Statistics, and Measurement
- Yet success on this small area provided the best predictor of passing and failing
- ***Why?***

Problematic Example #2

- This relates to reading tests, often known as tests of reading comprehension
- It is easier said than done to write difficult reading comprehension questions
- Vocabulary is a key source of difficult questions on verbally loaded questions
- So test developers often use vocabulary to increase differences

What is not tested in a paper-and-pencil test?

- Actual hands-on skills and performance
  - Must be assessed in other ways
  - Must be assessed over time and over situations
- Ethical issues

The Test Questions

- Who wrote them?
- What were the qualifications of the writers?
- Do they understand underlying theories and so on?
- How were they reviewed editorially?
- Were they pretested? (Go to item analysis section)

Item Analysis: How were the items pre-tested?

- On what sample?
- What analytic procedure?
- Are there multiple scales?
- How were the scales achieved?
  - Logically (by one or more people)
  - Empirically (by what procedure? Factor analysis?)
- What kinds of construct validity evidence can emerge from these analyses?

The Role of Consequences

- Consequences have always had a role in validation (Kane)
  - Intended consequences after all are criteria
- The traditional negative consequences were adverse impact on ethnic minority groups (known as adverse impact)—this is what most concerned Messick
- Debate over whether Messick's definition included consequences, although there is no doubt he saw it as important

The Role of Consequences

- Many others have argued validity should only be related to score interpretation (see Kane, 2006, p. 54)
- I share Popham's view that consequences are all always important; importance *per se* does not make them validity!
- We must differentiate the more important considerations such as UTILITY, to which validity is but a major part

Fairness

- Is the test fair to all demographic groups?
- Is looking at average scores a good way to determine fairness?
- Look at whether the relationship between test scores and criterion performance is equivalent across groups
- Look at differences on specific item performances across groups, equated for overall score levels.

Case #2

- Sally is a 17-year old who has a movement disorder. Throughout her education, she has been an average to somewhat below average student. She is in a motorized wheelchair. She cannot stand on her own. While she has complete use of her hands, her hands move rather slowly because she must concentrate completely on what she is doing. She has not been in special education, but has had accommodations in terms of classrooms, desks, physical education, and the like. She has a special interest in the use of computers, but not in programming. She hopes to go to secretarial school and become an office worker as she leaves high school with a diploma. The secretarial school boasts 100% employment of its graduates, but has a typing test as an entry requirement. Sally is willing to take the test, but does not wish to do so in a timed condition, as all other candidates are tested.

- ***As a professor at a local university, you are contacted and asked an opinion. Please provide it.***

Case #3

- You are an independent testing consultant working in the New York City area. One of your clients is a wealthy local school district. (Usually you help them in writing reports for the school board based on state test results.) They call you with an unusual problem. John Westminster Smith III, one of their students, has been a fine student throughout his 11 years of education. He has always been in normal classes. His parents have contacted the school district, however, due to the fact that he has been denied a 50% time extension (150% time) by the College Board on the SAT. The school district also shares with you a letter written by a local clinical psychologist who states that he believes that John as a learning disability and needs additional time. The letter was written last month.

- ***What would you recommend? What questions might you ask?***

Case #4

- Carol has graduated law school. She has *Multiple Sclerosis (*MS) and has received extended time on examinations, rest pauses, and freedom to stop the test at any time to receive medication, relax or use the washroom. State law forbids extension of time on the bar examination. She is considering suing to be able to take the test with extended time under the Americans with Disabilities Act. The position of the State Bar is that lawyers must be able to work under time pressure. They report that this is part of the purpose of the examination. She has not been able to find a lawyer to represent her and after consulting with professors at the law school, they refer her to you, a professor and testing expert with offices a couple of buildings away from the law school.

- ***What do you say to her? What kinds of documentation is she likely to need?***

Case #5

- The Law School Admission Council administers the *Law School Admission Test* (LSAT). Their test measures a variety of constructs, from logical thinking, reading comprehension, writing and general knowledge. All definitions of their constructs strongly involve the importance of speed. The Department of Justice believes that the test is too speeded and unnecessarily penalized individuals for whom speed of responding is a problem. They would like you to be a consultant supporting their case.

- ***What questions would you ask in deciding whether to work for them?***

Case #6

- As a professor of educational measurement at a university, you are active in the local psychological association and have become friendly with a number of psychologists who attend their monthly meetings. You have done some test construction consulting with an industrial psychologist who works for a local branch of a Fortune 500 company. He calls you with a problem that he is facing and reports that an individual who is blind has applied for a job which to receive, she would need to pass some tests. The tests were locally built. He reports that there are no versions of the test available for people who are blind and furthermore, the company does not have the resources to accommodate the work life of such a person.

- ***What would you advise this industrial psychologist?***

Some History

- Flagging and changes
- Testing's orientation toward disability (think of things that have really changed over time)
- ETS
- College Board

Hollenbeck: Test Alterations

- Defining what changes to testing are *appropriate* is key
- Differentiation of accommodations and modifications
  - Some use these terms interchangeably
  - Tindal et al. differentiated them
  - This issue may be the crux of legal hearings

Accommodations

- Remove construct-irrelevant variance
- Provide access to testing materials that would otherwise be missed
- 3 Attributes
  - Alters presentation or response mode to provide access
  - Does not change tested construct
  - "Differential access": score increases for those receiving the accommodation
    - Phillips asked how normal test takers would benefit if they received the accommodation
    - Reliability and validity of accommodation determination

Modifications

- The construct begin assessed is changed
- Involves a significant change in the test
  - From reading to having something read to the person
  - From a highly speeded test to extended time
- Do not interpret the same as "normal" scores
- Only alternative may be not testing

Requirements for Test Alterations to be Accommodations

- 1. Unchanged constructs
- 2. Individual need
- 3. Differential effects
- 4. Sameness of inference (really #1)

Unchanged Constructs

- Accommodation eliminates confounding factors, leaving construct to be assessed
- Example: eliminating reading from math tests
- Removing speed from tests of aptitude
- What do you think about these changes?

Individual Needs

- Necessity for assessment
  - Simply not possible?
  - Not possible to get a valid score?
    - Eliminates irrelevant factors/variance
  - Even with accommodations, not possible to get valid scores
- Competing interests: what must be done to help individual get a valid score (not optimal score)

Differential Effects

- Accommodations should differentially impact an affected group more than the unaffected group
  - Known as "differential boost"
  - Think of differential validity from the test bias literature
- Possible legal implications
- Sireci *et al.'s* point

Sameness of Inference

- Accommodations "work for those who need it and not for others"
- Validity requires that "any score inference from an accommodated test should be the same as for standard administration" (p. 399)
- Stakeholders need to have confidence
- No biases, no inflation, no deflation

Requirements of IDEA '97

- States determine appropriate accommodations
- High-stakes programs set criteria and acceptable accommodations
- Push to be inclusive in testing
  - For what reasons?
  - IEPs critical
  - IEPs operate like separate ships passing in the night

Teachers: Measurement Experts

- IDEA is heavily dependent upon IEPs
- Teachers are not measurement experts
  - Even special educators don't know a lot about assessment (depends on states/educational diagnosticians)
- Series of studies of teachers' knowledge
  - Possible research area
  - Inconsistent application of state policies by teachers (SpEd not better than GenEd)
  - Legal precedent in CA re AAMC determinations
  - "caution when relying on teacher judgments about test accommodations" (p. 401)

Accommodations & Validity

- Paradox:
  - Goal of accommodations is to ensure valid assessment
  - Accommodations may add extraneous variance
- Inconsistent application of accommodation rules=extraneous variance
- Relatively little research on accommodations
  - In recent years: labs, read-aloud accommodations

Categorizations of Disability: One Example

- Learning differences
- Emotional differences
- Sensory differences
- Physical differences

Settings for Testing: Physical Environment

- Can student work independently?
  - Prone to distractions, needs quiet, will bother others (e.g., ADHD)
  - Motivated to do well (time does not help)
- Familiarity with testing environment
  - Familiarity with teacher (issues?)
- Physical abilities, vision, hearing

Test Administration/Presentation

- Large-type, read-aloud, computers, calculators, pacing issues, language
- How does the student normally read?
  - In what manner (e.g., large-type)
  - How effectively (at grade level?)
- Can student follow directions?
- Does student normally use computers?
- How tested in the classroom?

Test Administration/Presentation Research Findings

- Large print: positive findings
- Read-aloud: mixed findings
- Pacing: student-controlled pacing significantly helps those with disabilities
- Computer presentation: paper & pencil preferred
- Language simplification: too few findings, but they showed no differences

Test Responses

- Extended time? (Biggest issue)/ story
  - Complex findings—Sireci's conclusions
  - Most findings with college admissions tests
- Ability to physically manipulate test materials?
- Ability to write? To type?
  - Marking in books—mixed results, few differences
  - Dictating to a scribe: strong positive findings
  - Word processing responses: mixed findings

Issues Involving Raters

- Raters know about accommodation of performance rating (e.g., word processing)
- Rater reliability is generally low
- Hard to find differences given same

Decision-Making Flowchart

- Excellent overview
- Recommendations
  - Choosing the right adjustment is difficult, "at best effective only for individual students" (p. 417)
  - Require qualifications to serve on decision-making bodies
  - More informed educators/testers
  - Operational rules for making alteration decisions
  - Follow model such as flow-chart

Alterations = Accommodations

- Changes in physical environment
- Test giver familiarity (some remaining issues)
- Marking in the test booklet
- Pacing (self-pacing)

- ***What makes an alteration or modification an acceptable accommodation?***

Alterations Perhaps Accommodations

- *Read-alouds*
- Computer-assisted testing
- Extended time

Helwig's Methodology for Creating Alternative Assessments

- Many LD students excluded from statewide testing
- IDEA required all students to participate
- Test accommodations are not sufficient
  - Modifications, including a change in content of the test
  - Don't measure the same construct (at least equivalently)

High Standards and LD

- We are sinking (or are we)?
- Virtually all states require testing (49)
- Many students do not meet state standards
  - Chapter cites states where more than ½ do not meet Basic standards
  - LD students often have poor language, writing & reading skills, poor memory, other deficiencies
  - Performance measures are also difficult for LD

Why look at LD students?

- High incidence
- Not studied much with statewide tests
- Did comparison with general education students

Their Task

- Modify a reading comprehension test
- Start by identifying tasks that are difficult for the population
- Identify components needing modification from literature
- Field test

Example Issues

- Long reading passages insurmountable
  - Vocabulary test as approximation for reading?
  - Spelling tests?
  - Others (Construct underrepresentation)
- Tests of components of reading
- Still attempt to address state goals

Issues of Field Testing

- Establishing a Measurement Standard
  - Match to the Standard Measure
  - Correlate with a common measure (seems like concurrent validity below)

- Establishing Concurrent Validity—establish match (r) to standard measure
- Establish Structural Validity (reliability, factor structure)—can do confirmatory factor analyses
- Validity/Generalizability across groups
- Overall goal: strong construct validation

Other Requirements of Alternative Measures

- Alternate forms for administration
- Quick scoring for feedback
- Generally aimed at low performing students

Some Individual Educational Plan Issue

- IEPs must include statements of test modifications
- Prior to 1998, 50% of children (with IEPs) were excused from large-scale assessments

Who Qualifies as a Student with a Disability

- 13 disabilities mentioned in IDEA
- 4 Categories account for 90%
  - Speech/language impairment
  - Serious emotional disturbance
  - Mental retardation
  - Serious Learning disabilities
- Different locations define all differently
- IDEA criteria used for admission to special education, along with need for special educational services

IEP Processes

IEP Process:

- Plan reflects each student and the services he/she needs
- Why would such children be denied testing?

Legal and Professional Standards

- Legal: Section 504 of Rehabilitation Act of 1973, IDEA, ADA
  - Guaranteed services
  - Standards for Educational and Psychological Testing (1985, 1999, 2014)

Thank You

BUROS
CENTER FOR TESTING