

Testing Issues and Concerns: An Introductory Presentation to the State Directors of Career and Technical Education

Kurt F. Geisinger, Ph.D.
Director, Buros Center for Testing
University of Nebraska-Lincoln
kgeisinger2@unl.edu

The Buros Center for Testing

- Rutgers, 1938; University of Nebraska, 1979
- To improve the science and practice of testing
- Review commercially available tests (BIMM)
 - Mental Measurements Yearbook
 - Tests in Print
- Consult and perform research on practices for proprietary testing programs (BIACO)
 - Education, admissions, credentialing, and employment
 - Projects with a number of states regarding K-12 and certification testing
 - Accredited testing programs
 - Evaluated NAEP

Kurt F. Geisinger, Director

- Built nursing examinations during graduate school
- Validated Veterinary Aptitude Test
- Taught/Directed Psychometrics program at Fordham U. for 15 years
- Built dozens of civil service tests for New York City (police, fire, sanitation, health services, etc.)
- Testified in about 2 dozen testing court cases
- Served on national GRE and SAT committees
- Edited or wrote 4 testing books

Goals of this Presentation

- To initiate a discussion of some testing topics for your consideration
- To remind and inform participants of the benefits and concerns with testing
- To help this group focus on what it hopes to accomplish with a national testing program
 - To identify primary and secondary purposes
 - To understand that there are costs associated with each possible purpose
- To begin to form a strategic plan that would move Career and Technical Education ahead in this regard

Purpose of a Test

- A clearly defined purpose for any test is critical
 - The purpose defines
 - How a test is constructed
 - How it is used
 - How it is evaluated
 - Sets parameters on how to improve it
- Needs to be determined in a manner that is particular to a specified population
- Tests can have multiple uses
 - Each may require different characteristics
 - Each needs to be evaluated

Types of Tests/Decisions

- Tests differentiated by uses
 - High stakes vs. low stakes tests
 - High stakes: admissions, licensure, graduation requirements, exit requirements
 - Low stakes: outcomes assessments, experimental tests
 - Mid-level stakes: class tests, placement tests, tests for credit
- Paper-and-pencil vs. Performance tests (where do computer-administered tests fall?)
- Norm-referenced vs. Criterion-referenced tests

How do we evaluate tests?

- Major Criteria:

- Reliability

- Stability over time
 - Stability across items supposed to measure the same thing

- Fairness

- Other aspects of impact or usefulness

- Quality of test construction

- Validity (by far the most important)

- Degree to which test accomplishes its goal
 - Degree to which test encompasses the field
 - Degree to which scores from test are meaningful
 - Imagine two tests drawn from the same pool
 - Test security is a factor

Reliability

- **Test-retest:** when a group takes the same test twice, one would expect their scores to be similar each time
- **Alternate-forms:** if we sample two sets of items from the item pool, scores from both should be similar
- **Internal Consistency:** does one's performance on different items composing a test seem to be similar and make sense?
- Reliability sets limits on validity

Fairness

Do different groups perform similarly? In a reasonable fashion?

If groups differ on their ultimate scores, do they do so in a meaningful manner?

Are there reasons for differential performance?

Typical groups: men/women, ethnic groups, racial groups, those with disabilities, language minorities

There is litigation potential when groups differ

Fairness of Computerized Tests

- If tests are randomly assembled, can some complain that they received a randomly too difficult test?
 - That some received too many items in one area of the content and not enough of another?
 - That some are not familiar with the computer?
 - That their test differs in some meaningful way from that of another test taker?
 - That other students received some of the items in advance from their colleagues? Their teachers?
- Some students with disabilities will rightly demand accommodations under ADA

Aspects of the Impact of Testing

■ Example concerns:

- How difficult is the test to administer?
- How do students feel about the test?
- Does the press get involved? How?
- Do “coaching schools” such as Kaplan provide help to students and is that help useful?
- Are students practiced in the manner of testing?
- Is the level of English language required by test items consistent with that of the students?

Quality of Test Construction

- Are items fair to students? (e.g., are they tricky?)
- Do they follow good item writing principles?
- Do they test students in terms of the deeper knowledge expected of them or just memory?
- Do they represent the field(s) appropriately?
- Is the test long enough to provide useful scores?
 - Are sub-tests similarly long enough?

Item Writing

- Good item:
- For what is George Washington best known?
 - Throwing quarter across the Potomac
 - Moving the Capitol to Washington, Dc
 - Sleeping in dozens of places
 - Being the first president of the country
- Bad item:
- George Washington:
 - Threw quarter across the Potomac
 - Moved the capital to Washington, DC
 - Slept in dozens of places
 - Was first president of the US

Validity: Some General Concepts

- Defines the extent to which the goals of the test are being appropriately met
 - Does test do what it was intended to do?
- Lots of ways to prove a test is valid (for a specific use)
- Tests are not universally valid
- The higher the stakes for which a test is used, the more necessary the validation and the higher the validity should be
- Why do students complain about the inappropriateness of tests?

Content Validity

- This is one kind of validity
- Content validity deals with the degree to which a test represents the areas it should cover
- For example, if a test needs to cover 16 areas, would you be happy if it covered 12?
 - Would you be happy if it covered all of them, but not very well?
- Is test to be used to make individual decisions? For programmatic decisions? (Different reliability and validity required for each)

So What Would be the Purpose of the Test Item Bank for Career & Technical Education?

■ For Students

- To assess entry into programs
- To assess worthiness for federal aid
- To assess learning in specific courses or classes
- To determine whether a student should receive credit for a class or have its requirement waived
- To assess learning in a program
- To assess competence for entry into a profession
- Others?

So What Would be the Purpose of the Test Item Bank?

- For Programs
 - For formative feedback to help shape the curriculum
 - For formative feedback to help teachers meet student needs
 - For summative feedback on program quality, both internally and externally
 - For accreditation purposes
 - For state and federal accountability purposes
- Would individual tests work better?

Differences Among Purposes

- Differences among these purposes impact:
 - Test development
 - The very nature of the test
 - Validity needs
 - Test security needs (and who cheats!)
 - How often the test items need revision and updating
 - Frequency with which tests can be administered
 - Costs! (The higher the stakes, the higher the costs)

Who decides what the nature of the test is?

- Process issues
- Funding issues
- Who will do the work? (benefits of multiple partners)
- Can your group justify the use that you wish to make of the test(s)?

My recommendations

- Enumerate the needs justifying the test bank
- Survey the extent of each of the needs
- Discuss the different needs
- Identify the priority needs for the tests
- Work with testing expert(s) to plan for the tests
- Consider delivery systems and services
- Develop tests consistent with their purpose(s)
- Validate the uses to which the test(s) is/are put
- Make sure that test security is commensurate with the uses and stakes

Thank you for your kind attention?

- Hopefully I have been answering your questions right along, but this is the time for doing so!
- If you have questions or comments:
Kurt F. Geisinger at the University of
Nebraska-Lincoln
kgeisinger2@unl.edu