Written Multiple-Choice Job Knowledge Tests: Pros, Cons, Misunderstandings, & Admonitions

2011 IPAC Conference – Capital Ideas for Assessment

Washington, D.C.

Kyle E. Brink
St. John Fisher College

Jeffrey L. Crenshaw
Centrus Personnel Solutions

Written Multiple-Choice Job Knowledge Test

Pros

- Cost effective
- Reliable
- Valid
- Objective
- Fair/Unbiased

Cons

- Expensive
- Unreliable
- Invalid
- Subjective
- Unfair

Written Multiple-Choice Job Knowledge Test

- Definition: Measures knowledge required for job
- Typical characteristics
  - Assess job knowledge (often technical)
    - Sometimes attempt to assess skills, abilities, and behaviors
  - 100 multiple-choice items
  - Based on source materials
  - Closed book
  - Cut score = 70
- Commonly used in civil service
  - Especially common for police & fire promotions
- Rarely used in business or private sector

Written Multiple-Choice Job Knowledge Test

- Critique qualities of multiple-choice job-knowledge tests:
  - Objectivity
  - Reliability
  - Validity
  - Cut-scores
  - Ranking
  - Adverse impact
  - Test security
  - Cost
- Suggestions for practice

Objectivity

- Objectively scored: test-taker receives same score regardless of who scores it
- The concern: rater bias and favoritism
  - Subjectively scored tests could allow for real or perceived favoritism
- The misguided solution:
  - Instead of fixing the problem of favoritism, many organizations eliminate subjectively scored tests

Objectivity: The Reality

- Objectively scored tests are subjective too
  - Subjective decisions related to:
    - Job analysis, choosing source materials, choosing item content, writing items, choosing distractors, writing distractors, choosing final items
- Objectively scored ≠ reliable
  - Reliability is an important psychometric property that impacts validity
- Subjectively scored ≠ bias or unreliable test
  - Eliminate source of bias instead of changing to different test

Reliability: Objective vs. Subjective

- Subjectively scored tests can be reliable
- Objectively scored tests can be unreliable
- Arthur, Edwards & Barrett (2002) compared
  - Objectively scored multiple choice vs.
  - Subjectively scored open-ended responses
    - "Subjective" test was more reliable
    - Subjectively scored test resulted in smaller race differences and less adverse impact
      - "In some instances, substantially so"
      - In some instances, Black candidates passed at a higher rate than White candidates

Estimating Reliability

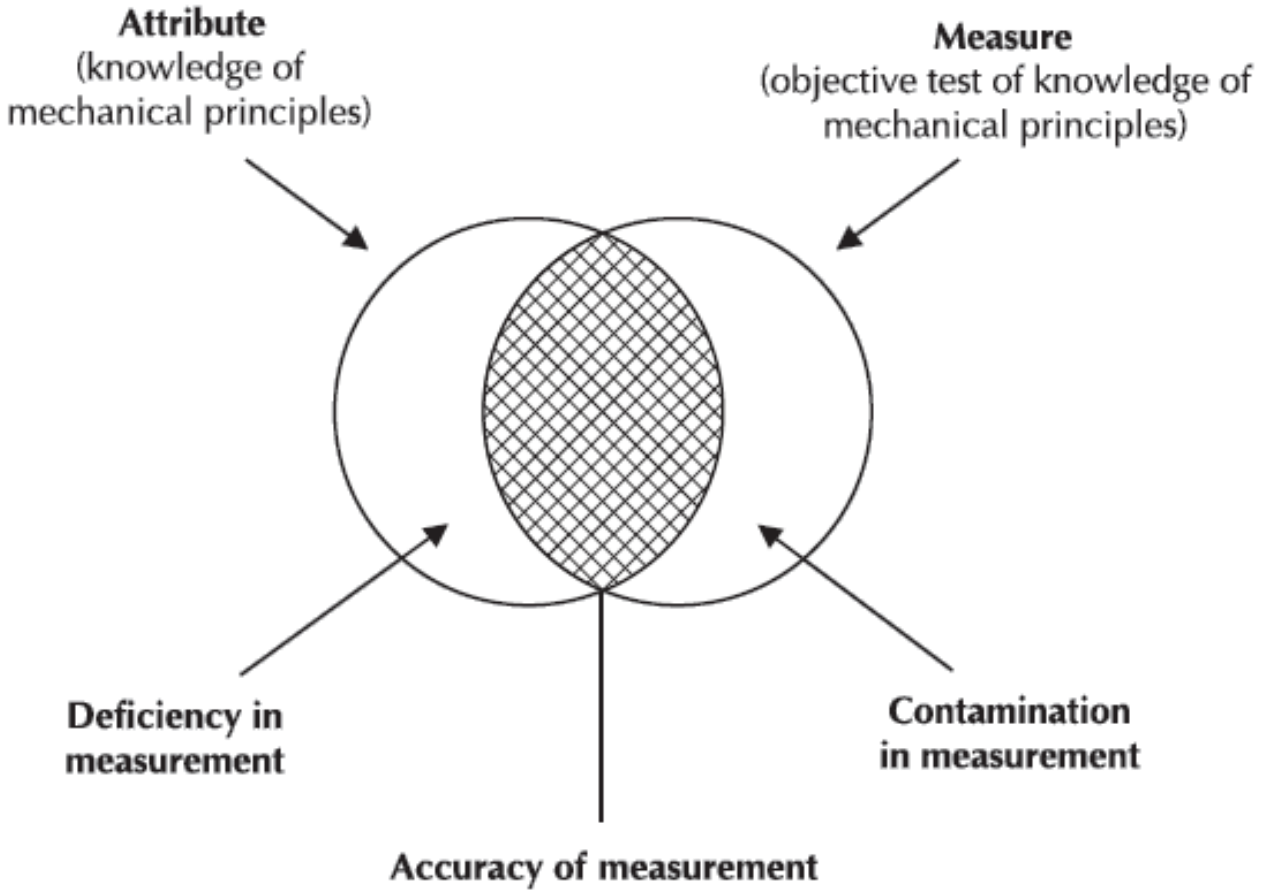
- Test-retest is not typically possible
- Coefficient alpha
  - Often low
    - Multidimensional
    - Lack of variance (especially if many easy/hard items)
  - Estimate by dimension if appropriate
    - Estimate overall reliability with the reliability for a linear composite formula (Nunnally & Bernstein, 1994)

Validity

- What do we mean by "valid" test?
  - Accuracy of measurement
    - Degree to which a test truly measures the attribute it is intended to measure
  - Accuracy of prediction
    - Extent to which performance on the test is associated with performance on the job

Source: Heneman, Judge & Kammeyer-Mueller (2012)

Accuracy of Measurement



Source: Heneman, Judge & Kammeyer-Mueller (2012)

Accuracy vs. Contamination

- Which are important and needed upon entry in the job analysis?
  - Job knowledge
  - Ability to learn a body of information
  - Reading comprehension
  - Test-taking ability
  - Motivation to study
  - Ability to memorize
  - Leisure time
  - Disposable income

Content Validity

- What do we mean by “content valid” test?
  - Representative of the job
    - Test considered more content valid to the extent that it assesses a more representative sample of the job
  - Resemble the job
    - Test considered more content valid to the extent that KSAs that are required by format of the test are the same as those KSAs required to perform the tasks required on the job (Goldstein, Zedeck & Schneider, 1993)
  - Both test format and test content are important

Representative of Job

- SIOP Principles (p. 25)
  - *"Evidence for validity based on content rests on demonstrating that the selection procedure adequately samples and is linked to the important work behaviors, activities, and/or worker KSAOs (knowledge, skills, abilities, and other characteristics) defined by the analysis of work"*
- Uniform Guidelines (Section 14C[4])
  - A job analysis is required for content validity studies
  - *"To demonstrate the content validity of a selection procedure, a user should show that the behavior(s) demonstrated in the selection procedure are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job"*

Representative of Job

- Uniform Guidelines
  - ***79***. *Q. What is required to show the content validity of a test of a job knowledge?*

    A. There must be a defined, well recognized body of information, and knowledge of the information must be prerequisite to performance of the required work behaviors. The work behavior(s) to which each knowledge is related should be identified on an item-by-item basis. The test should fairly sample the information that is actually used by the employee on the job, so that the level of difficulty of the test items should correspond to the level of difficulty of the knowledge as used in the work behavior. See Section 14C(1) and (4).

Representative of Job

- Knowledge ≠ responsibilities, skills, or abilities
  - Knowing how to perform an activity may be a prerequisite to performing that behavior, but knowledge alone is not sufficient
    - Just because an individual possesses the knowledge doesn't mean he/she can perform the responsibility or possesses the skills or abilities that are required to perform the responsibility (Goldstein, 1993, p. 62)
  - Assessment of responsibilities, skills, and abilities via a "job knowledge" test cannot be valid

Representative of Job

- Using a job knowledge test alone results in a selection process that is severely deficient in representing actual job responsibilities
  - Does not adequately represent the job (and cannot be content valid) because it only assesses knowledge and does not assess responsibilities/work behaviors, skills, or abilities
  - A closed book test is even more deficient because it is not assessing knowledge that doesn't need to be memorized
- Any selection procedure that includes more constructs than just job knowledge would be more content valid than a selection procedure comprised of only a job knowledge test

Content Validity

- What do we mean by “content valid” test?
  - ✓ Representative of the job
  - – Resemble the job

Resemble the job

- SIOP Principles (p. 24)
  - *"The more a selection procedure has fidelity to exact job components, the more likely it is that the content-based evidence will be demonstrated"*
  - *"Generally, the more closely a selection procedure replicates a work behavior, the more accurate the content-based inference"*

Resemble the job

- Uniform Guidelines (Section 14 C (4))
  - *"The closer the content and the context of the selection procedure are to work samples or work behaviors, the stronger is the basis for showing content validity. As the content of the selection procedure less resembles a work behavior, or the setting and manner of the administration of the selection procedure less resemble the work situation, or the result less resembles a work product, the less likely the selection procedure is to be content valid, and the greater the need for other evidence of validity."*

Resemble the job

- Uniform Guidelines
  - *78. Q. What is required to show the content validity of a paper-and-pencil test that is intended to approximate work behaviors?*

    A. Where a test is intended to replicate a work behavior, content validity is established by a demonstration of the similarities between the test and the job with respect to behaviors, products, and the surrounding environmental conditions. Section 14B[4].

    Paper-and-pencil tests which are intended to replicate a work behavior are most likely to be appropriate where work behaviors are performed in paper and pencil form (e.g., editing and bookkeeping). Paper-and-pencil test of effectiveness in interpersonal relations (e.g., sales or supervision), or of physical activities (e.g., automobile repair) or ability to function properly under danger (e.g., firefighters) generally are not close enough approximations of work behaviors to show content validity.

Resemble the job

- Most jobs do not require incumbents to make written, multiple-choice decisions
- Almost any alternative method of assessment would more closely resemble the job than answering multiple-choice questions
- Using a written, multiple-choice job knowledge test alone cannot be supported based on content validity

Validity

- Criterion-related validity
  - Rarely done with job-knowledge tests
    - Typically attempt to minimize exposure and only use the test once due to test security
- Transportability/validity generalization
  - Assumes same/similar job and test
  - Rarely done with job knowledge test
- Most recent meta-analysis estimate – Hunter & Hunter (1984) – Job Knowledge

Setting Cut Scores

- Critical Score
  - Minimum threshold; minimally competent vs. incompetent

  Vs.
- Cut Score
  - Pass/fail point based on any number of factors
    - Number of positions, number of applicants, adverse impact, etc.
- What is the justification for 70% pass-fail cut?

Setting Cut Scores

- Requirements for Cut Scores as typically used:
  - A reasonable rationale is necessary
  - Set at the minimum level necessary for competent performance as a newly promoted employee
  - Set at the level that will protect the public from the harm of low performance
- Use SME driven approach (e.g., Angoff Method)
  - SMEs should be chosen based on expertise
  - Most advise using 7-15 SMEs (balance test security considerations)
  - Training is critical - cover purpose (to develop cut score), the tools/ information/ test available to inform judgment, the exact judgment task they will be asked to do
- The established cut-off is sometimes adjusted (usually downward) after the exam is given based on the standard error to account for the reliability/precision of the measure

Ranking

- Adverse impact is typically greater when using a job knowledge test on a ranking basis versus using it on a pass/fail basis
  - If candidates are hired based on rank order and adverse impact occurs, the *Uniform Guidelines* (Section 5G) require sufficient validity evidence to support the use of a test based on ranking purposes

Ranking

- A test that assesses only knowledge does not provide a strong basis for ranking
  - Uniform Guidelines Questions & Answers No. 62
    - *"Where the content and context of the selection procedure are unlike those of the job, as, for example, in many paper-and-pencil job knowledge tests, it is difficult to infer an association between levels of performance on the procedure and on the job."*
  - APA Standards (p. 161)
    - *"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."*

Ranking

- Additional concern with ranking based solely on a multiple-choice job-knowledge test:
  - Job knowledge tests often lack sufficient reliability and variability in scores to support meaningful differences in test scores
  - Even though candidates may obtain different scores, there is little confidence that those differences reflect true differences in job-related abilities
- Use as pass/fail
  - Rank on other test components or on full battery of tests as appropriate

Adverse Impact/Group Differences

- Not clearly known; surprisingly little research
  - Only one meta-analysis estimate (Roth, Huffcutt & Bobko, 2003)
    - Job knowledge as criterion variable; NOT as predictor
    - *d* = .48 (favoring White over Black)
- Race differences are expected to be relatively large
  - Written format
  - Similar to and correlated with cognitive ability

Adverse Impact/Group Differences

- Edwards & Arthur (2007) compared
  - Objectively scored multiple choice vs.
  - Subjectively scored open-ended responses
    - Subjectively score test resulted in 39% reduction in subgroup differences
    - Black test-takers had more favorable perceptions on the subjectively scored test
      - Higher perceived job-relatedness
      - Higher perceived fairness
      - Higher test-taking motivation

Adverse Impact/Group Differences

- Incorporate additional constructs
  - Ployhart and Holtz (2008)
    - The most effective strategy for reducing subgroup differences with no validity tradeoff is to assess the full range of knowledge, skills, abilities, and other characteristics (KSAOs).
  - APA Standards
    - *"Success in virtually all real-world endeavors requires multiple skills and abilities"* (p. 79)
    - *"Issues of fairness may arise in the choice of which factors are measured"* (p. 80)

Test Security

- Sharing answers (or scoring keys) with test-takers:
  - Greatly increases cost of testing
  - Is not done in any other industry
  - No professional guidelines mandate or even encourage providing scoring keys to test takers
- Why do we insist on sharing answers with candidates?

Test Security: APA Standards

- Entire chapters on
  - *The Rights and Responsibilities of Test Takers*
  - *The Responsibilities of Test Users*
  - Nowhere does it mention that test takers have the right to see the scoring key or that test users are required (or even advised) to provide test takers with the scoring key
- Test users have the responsibility to protect security of tests
- Developers/users should monitor for scoring errors and ensure procedures are in place to prevent scoring errors and rescore if requested
- Should provide feedback, but need rigorous protection of test security

Test Security: SIOP Principles

- *"Public disclosure of the content and scoring of most selection procedures should be recognized as a potentially serious threat to their reliability, validity, and subsequent use. All data should be retained at a level of security that permits access only for those with a need to know."* (p. 44)
- *"Scoring keys should not be included in technical reports or administration manuals and should be made available only to persons who score or scale responses."* (p. 56)
- *"The researcher should include information on how to provide feedback to candidates, if such feedback is feasible or appropriate. Feedback…should not violate the security of the test or its scoring."* (p. 57)

Cost

- Relatively cheap to administer
  - Group administration
- Moderate expensive to develop properly
- Typically only used once

Suggestions for Practice

- Objectivity:
  - Choose measure that most effectively assesses what is important based on the job analysis
  - Don't choose a measure because it is objectively scored
- Reliability
  - Difficult to accurately estimate
    - Inflated because of number of items
    - Deflated because of heterogeneity of items

Suggestions for Practice

- Validity
  - Never use job knowledge test alone
    - It is not representative of the job
    - It does not resemble the job
  - Use the job knowledge test for what it is designed to do - assess job knowledge
  - Only measure knowledge identified in job analysis
  - Choose the test that best assesses the job-related construct being assessed
    - There are other ways of measuring knowledge
    - Other methods are better at application of knowledge and at assessing behaviors, skills, abilities
    - Assess knowledge that does not have to be memorized…in an appropriate manner

Suggestions for Practice

- Cut
  - Do not automatically use a cut of 70%
  - Use professionally accepted methods to establish meaningful cut scores - minimum level necessary for competent performance
- Ranking
  - Use as pass/fail instead of ranking
- Adverse impact
  - It will occur on traditional multiple-choice job knowledge test
  - Therefore, heed all prior suggestions and use wisely
- Test security
  - Do not turn over more information than necessary

Questions?

References & Source Materials

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.
- Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology, 55*, 985-1008.
- Barrett, R. S. (1996). *Fair employment strategies in human resource management*. Westport, CT: Quorum Books.
- Barrett, R. S. (1998). *Challenging the myths of fair employment practices.* Westport, CT: Quorum Books.
- Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education, 8*, 99-109.
- Biddle, D. (2006). *Adverse Impact and Test Validation* (2nd ed.). Hampshire, England : Gower Publishing Limited.
- Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1-24.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

References & Source Materials

- Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8*, 360-362.
- Edwards, B. D., & Arthur, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*, 794-801.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures.* Federal Register, 43(166), 38290-39315.
- Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of Treasury (1979). *Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures.* Federal Register, 44, 11996.
- Gatewood, R. D., Feild, H. S., & Barrick, M. (2008). *Human resource selection* (6th ed.). Mason, OH: South-Western.
- Goldstein, I. L. (1993). *Training in organizations.* Pacific Grove, CA: Brooks-Cole.
- Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 3-34). San Francisco, CA: Jossey-Bass.

References & Source Materials

- Heneman, H. G., III, Judge, T. A., & Kammeyer-Mueller, J. D. (2012). *Staffing organizations* (7th ed.). Middleton, WI: Mendota House.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.
- Maurer, T. J. (2005). Distinguishing cutoff from critical scores in personnel testing. *Consulting Psychology Journal: Practice and Research, 57*, 153-162.
- Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Phillips, J. M., & Gully, S. M. (2012). *Strategic staffing* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. Personnel Psychology, 61, 153-172.
- Roth, P. L., Huffcutt, A. I., & Bobko, P. (1993). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*, 694-706.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.