Best Practices in Examinations: A Brief Look

Prepared for CISRO

Edwin L. Weinstein, Ph.D., C.Psych.
The Brondesbury Group

27 April 2010

What Makes a ‘Best Practice’

- Reducing or eliminating “Construct Irrelevant Variation” (CIV) is the basis of ‘best practices’.

- The ‘construct’ is whatever you are trying to measure. It SHOULD be the ONLY thing the exam measures.

- CIV are differences in test scores due to causes other than construct-related abilities of the examinee. Main sources are:
  - Test items
  - Test administration conditions
  - Test scoring including the setting of pass marks.

Best Practices in 12 areas

1. Practice analysis
2. Test specifications
3. Item development
4. Test production & distribution
5. Exam administration
6. Setting performance standards
7. Exam scoring
8. Score & decision reliability
9. Score equating & scaling
10. Score & decision validity
11. Score & decision reporting
12. Documentation

Ten Best Practices – Item Development

1. Item writers should reflect a mix of specialties plus other relevant demographics like gender.

2. Items must be authored by people who have had explicit training in item writing.

3. Items should be independently reviewed before editing. The author should not be a reviewer.

4. The reviewer should assess without prior knowledge:
   - Content domain
   - Bloom level
   - Sources of CIV including linguistic complexity & tricky wording.

Ten Best Practices – Item Development

5. The test item editor has two roles:
   - Screen for potential sources of CIV
   - Check for “15 fatal flaws” for items

6. All test items should be field tested on a pilot group of students prior to being counted on an exam.

Excerpted from “**MCQ: The 15 Fatal Flaws**”, Ray Talke, Minds in Action, Inc. NOCA Academy Presentation, April 28, 2009.

Improving Your Test Items

1. **Avoid Unfocused Item Stems**
2. Assess the Proper Cognitive Level
3. **Avoid Negative Item Stems**
4. **Use Complex Multiple-Choice Test Items with Discretion**
5. **Ensure Grammatical Congruence Between Item Stem and Answer Options**
6. Avoid Subjectivity
7. Avoid Qualifiers and Specific Determiners *(e.g., typically, always, generally, etc.)*
8. Ensure that the Content of Test Items is Current
9. Avoid Humour
10. Avoid Stereotypical Descriptions
11. Avoid Analogies, Metaphors, Colloquialisms and Regional Expressions
12. Ensure that Distracters are Plausible (to Unqualified), yet Clearly Incorrect (to the Qualified)
13. **Make all Answer Options Homogenous (Parallel)**
14. Use Jargon and Acronyms Only When Universally Recognized
15. **Avoid Overlapping Answer Options**

Some Examples of Fatal Flaws

- Unfocused Item Stem
  - Which of the following is true?
  - Underwriting is __________ .
- Negative item stems
  - Which of the following is not a feature of group disability policies?
  - Mutual funds include all of the following benefits except:
- Grammatical incongruence between item stem & answers
  - Publicly-traded stocks can be bought through a ... ?
    a. Insurance agent
    b. Stockbroker
    c. Mutual fund advisors
    d. Banker

Some Examples of Fatal Flaws

- Complex multiple choice items (e.g. multiple answer-multiple response)
  - Which of the following investment products pay interest?
    1. GIC
    2. Savings account
    3. Canada savings bond
    4. Preferred stocks

    a. 1,2 and 3
    b. 1, 3 and 4
    c. 1, 2 and 4
    d. 2, 3 and 4
- Answer options are not parallel
  - What is the main advantage of universal life compared to whole life?
    a. Permanent coverage
    b. Low cost
    c. You can re-balance investment & insurance coverage
    d. Automatic renewal
- Overlapping answer options
  - Which are features of a Registered Retirement Savings Plan
    a. Tax deferral, portability, broad range of investments
    b. Tax deferral, CDIC protection, portability
    c. Broad range of investment, portability, conversion to annuity
    d. Portability, CDIC protection, conversion to annuity

Ten Best Practices

7. Ongoing independent review of item statistics
   - Set parameters for item revise and/or drop
   - Criteria for dropping items with ‘competing distracters’
   - Criteria for dropping items with low success rate (e.g., p<.20)
8. Eliminate sources of cultural bias
   - Language structures incidental to content assessed
   - Other sources identified by credible research (e.g., timing)
9. Eliminate sources of physical discomfort or distraction
10. Reports regarding exam pass rates, candidate demographics, exams administered and exam site usage should be made available to the public, producer community, and industry.

Principles of Best Practice

- Independent judgment at every step.
- Multiple checks and balances.
- Regular evidence-based review of item suitability (pilot test and post-test item analysis).
- Eliminate construct-irrelevant variation
  - Clear language
  - No fatal MCQ flaws
  - Remove unnecessary physical distractions
  - Reduce/eliminate cultural and gender bias

References: Best Practices

- Standards for Educational and Psychological Testing (APA, CPA, AERA, NCME)
- S. Downing & T. Haladyna, Handbook of Test Development. New York: Routledge, 2006.
- National Association of Insurance Commissioners (NAIC), State Licensing Handbook, Ch.8: Testing Programs, 2008.
- R. Hambleton, Technical Guidelines for Evaluating Credentialing Exams, Public Accountants Council, 2007.

Thank You

The Brondesbury Group