

Assessment with Multiple-Choice Questions in Medical Education: Arguments for Selected-Response Formats



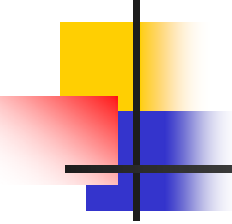
Congreso Nacional De Educacion Medica
Puebla, Mexico
11 January, 2007

Steven M. Downing, PhD
Department of Medical Education
University of Illinois at Chicago
sdowning@uic.edu



Arguments for SR

- Arguments based on science
 - Validity evidence, objective observation, psychometric and test development theory
- Arguments based on feasibility
 - Efficiency, logistic feasibility, speed, cost effectiveness



“Any aspect of cognitive educational achievement can be tested by means of either the multiple-choice or the true-false form.”

Robert L. Ebel, p. 103

(Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.)



Science – Content-related validity evidence

- Broad sample of content
 - 40-50 MCQs vs. 20-30 CR per hour
 - Reduces Content Underrepresentation (CU) threat to validity
- Systematic scientific evidence related to content sampled



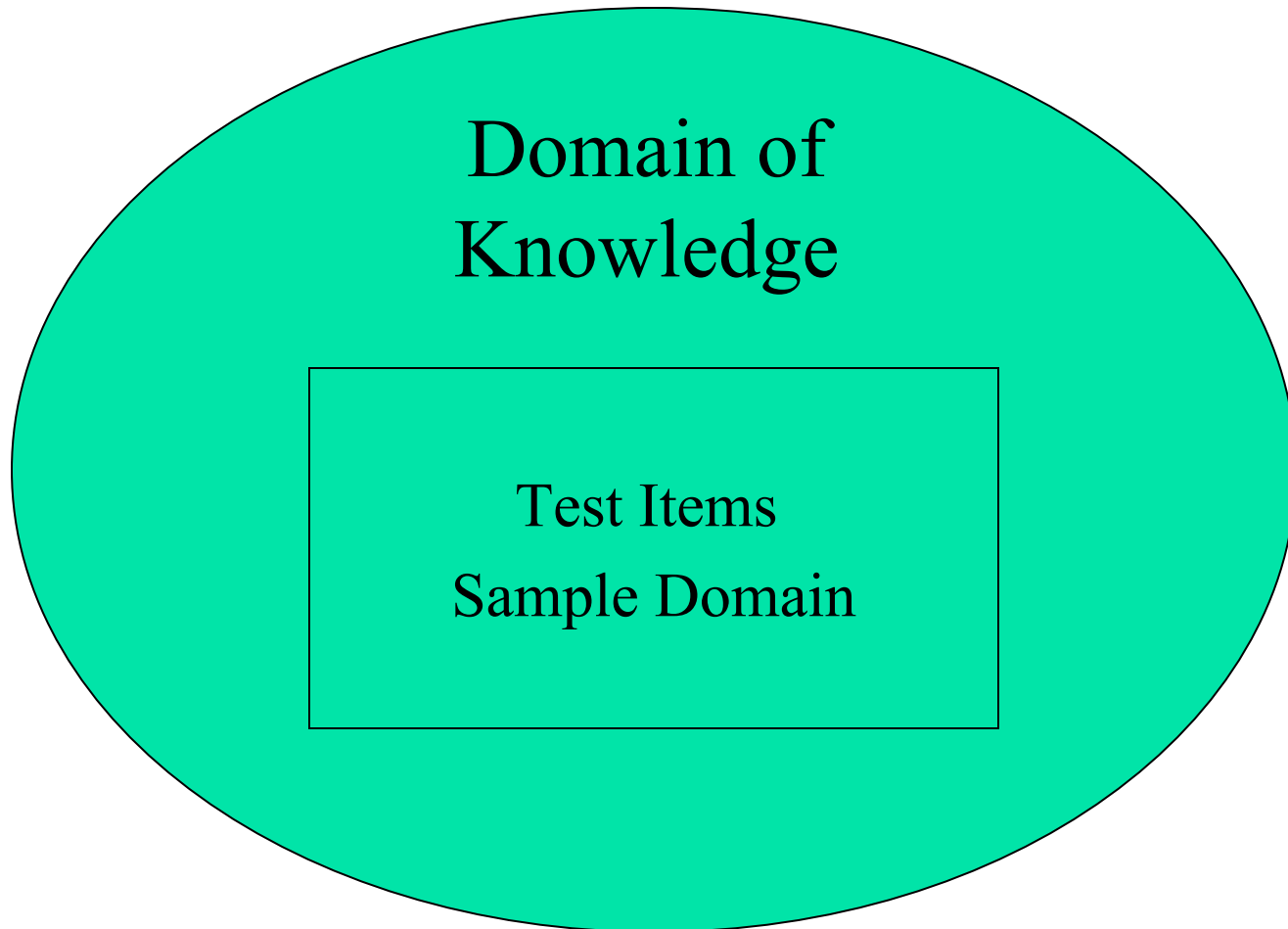
Science – Content-related validity evidence

- Systematic plan to sample content
 - Sampled test content related to entire content domain
 - Well established methods
- Test specifications – Blueprint
 - Item sample to total population
 - Inferences to domain



Knowledge Construct:

Inferences from Sample to Domain



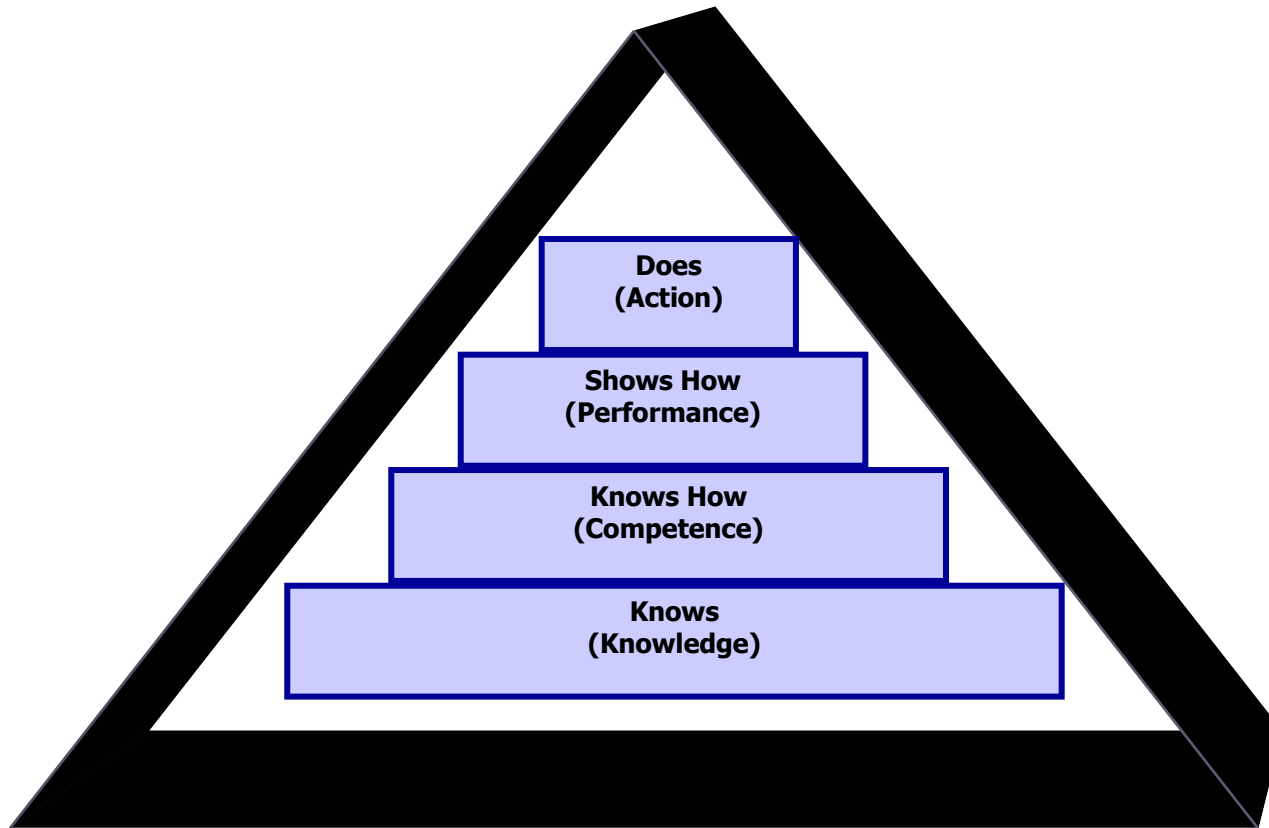


Science – Content-related validity evidence

- Items can test higher levels of cognitive knowledge
- More than recall and recognition of facts
 - Also, understanding, application, problem solving, judgment, evaluation and synthesis
- “Knows” and “Knows How” level of the Miller Pyramid



Miller's Pyramid



Miller, G. The assessment of clinical skills/competence/performance. Academic Medicine, 65(suppl): S63-67, 1990.



Factual Recall

Which of the following is a known cause of pancreatitis?

- a. Azathioprine
- b. Erythromycin
- c. Metoprolol
- d. Diltiazem
- e. Theophylline



Application to clinical setting

A 58-year-old man is admitted to the hospital with fever and chills. The patient has a history of angioedema and laryngospasm within 24 hours after receiving ampicillin. On day one of his hospitalization two sets of admission blood cultures show growth of gram positive cocci in clusters. What is the most appropriate empiric antibiotic therapy?

- a. Vancomycin (Vancocin)
- b. Imipenem/cilastatin (Primaxin)
- c. Cefazolin (Ancef)
- d. Clindamycin (Cleocin)
- e. Azithromycin (Zithromax)



Science – Objectivity

- Reproducibility
- Objective Scoring –
 - Objectively reproducible scoring rules
 - Rescore -- 100% agreement!
- Content expert agreement
 - Best, most correct answer
 - Least correct, wrong answers
- Defensibility of scores
- Enhances validity evidence

Science – Correlational Evidence



- What is the relationship between SR and CR scores?
- Meta-analysis¹: 29 studies with 56 correlations
 - Disattenuated correlations: SR - CR
 - Correlations estimated for perfectly reliable tests

¹ Rodriguez, M.C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. J. of Educational Measurement, 40: 163-184.



Science – Correlational Evidence

	Fixed Effects	Random Effects
Stem-Equivalent	0.92	0.95
Non Stem-Equivalent	0.85	0.86

Rodriguez, M.C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. J. of Educational Measurement, 40: 163-184.



Science – Psychometric Theory

- 100 years of development, research
- Examples –
 - Validity theory
 - Methods to quality control tests
 - Item analysis
 - Statistical analyses each item, each option
 - Routine summary test score analyses
 - Research-based methods for statistical equating of test scores



Science – Item Writing Theory

- Evidence-based principles
- ~100 years of development, research
- Many variants on MCQs
 - E.G., Extended matching
- Item writing challenge –
 - Trainable task
 - Specialized skills: write, review, edit items



Science – Test Development Theory

- Guided by requirements for validity evidence
 - MCQs maximize potential validity evidence
- Some theory, much research, lots of tradition



Feasibility – Efficiency of MCQs

- Cost effectiveness
 - MCQs more cost effective than CRs
- Most efficient for large groups of examinees
 - Upfront time to write, review, edit MCQs vs. time for two independent readers to score CRs
- Time efficiency
 - Maximum information per testing hour
 - Fast scoring
- Maximizes instructional vs. assessment time



Feasibility – Efficiency of MCQs

- Ease of test administration
- Accurate and reproducible scoring
 - Easily quality-controlled
- Ease of item banking –
 - Reuse: Secure storage and retrieval of effective MCQs
 - MCQs not as easily memorized as CRs



Feasibility – Efficiency of MCQs

- Establish passing scores readily
 - Absolute and/or relative methods
- Report scores to examinees quickly
- Provide useful feedback via subscores
 - Instructional value of immediate feedback
 - Profiles of student strengths-weaknesses



Summary - Science

- Strong content-related validity evidence
 - Representative sample of universe/domain
- Scoring objectivity
 - Reliable scoring
 - Independent replication without subjectivity
- Meta-analysis: SR – CR correlation high
 - R^2 : 72% to 90% variance in common
- Psychometric theory
- Long research and development history
 - Theory and practice
 - Evidence-based principles of item writing



Summary – Feasibility

- MCQs efficient and cost effective
 - Ease of administration, scoring, reporting
 - Greater information per unit of testing time
 - Fast feedback and score reporting
 - Ease of maintaining secure item pools for item reuse





Threats to Validity

- Two Major Sources of Validity Threats (Messick, 1989)
 - Content Underrepresentation (CU)
 - Construct-irrelevant variance (CIV)



CU: Content Underrepresentation

- Non-representative sample
 - Test fails to adequately sample population
 - Incorrect inferences to domain possible
- Examples
 - Too few essays (SR), oral prompts, MCQs, or OSCE cases to reliably sample domain



CIV: Construct-Irrelevant Variance

- Reliable measure of unintended construct
 - Good measure of an irrelevant construct
 - Anatomy essay test which measures writing skill more than anatomy
 - Written psychiatry test which measures reading proficiency
 - Internal Medicine performance test more associated with personality than patient communication competence
 - Oral exam in Pathology which is a better measure of student “stage presence” than understanding of path
 - Variable that interferes with intended interpretation or test score use



Application, calculation

How many days remain in the reign of George W. Bush?

- A. 800
- B. 753
- C. 750
- D. 739
- E. 722
- F. Far too many