Trusting Your Test Results: Building and Revising Multiple-Choice Examinations

Doug Franzen, MD, MEd
Monica M. Cuddy, MA
Jonathan S. Ilgen, MD, MCR

The Challenge

From informal assessments to high-stakes licensing examinations, multiple-choice questions (MCQs) are the most common method of assessing medical knowledge. MCQ-based examinations can be used to broadly sample knowledge domains, require fewer resources to administer and score than other assessment formats, and tend to produce consistent scores. To create examinations with scores that accurately support their intended interpretation and use in a particular setting, examination writers must clearly define what the test is intended to measure (the construct). Writers must also pay careful attention to how content is sampled, how questions are constructed, and how questions perform in their unique testing contexts.[1–3] This Rip Out provides guidance for test developers to ensure that scores from MCQ examinations fit their intended purpose.

Rip Out Action Items

Program directors should:

1. Be intentional when designing multiple-choice question assessments. Develop a blueprint and follow it. Pilot your items.
2. Collect and analyze performance data to regularly assess your assessments.
3. Overhaul poorly performing questions and examinations, particularly for high-stakes assessments, just as you would for any other curricular element. This is impossible without collecting and analyzing data on your test.
4. Accumulate validity evidence to support the intended interpretation and use of test scores, especially when using them to make high-stakes decisions.

What Is Known

Before writing MCQ "items," test developers should create a blueprint that outlines (1) the material to be covered by the test; (2) how that material is categorized into different content areas; (3) how those content areas link to the construct of interest; and (4) how much each content area should contribute to the overall examination.[3,4] Broad sampling of content more accurately captures the construct being tested, and longer tests increase the reliability of test scores. High-quality MCQ examinations should produce scores that reflect meaningful variation in the construct(s) being measured. To do so, item writers should consider how individual MCQs perform and how a collection of MCQs perform together.

Many factors can adversely affect the performance of items, including negative wording of items, implausible response options, grammatical errors, and cuing, such as when the correct answer can be inferred from the wording in the stem (question) or the structure of the answer choices.[3,5] Piloting items helps ensure that examinees understand what is being asked and that items perform as expected.[3] Item analyses, such as measuring item difficulty (the degree of challenge posed by the item) and item discrimination (the degree to which items differentiate between high- and low-performing examinees), can help assess item performance.[3,6] Test reliability (the degree to which the test produces consistent scores) can be measured a few different ways: test-retest performance, parallel forms, and internal consistency.[2,3,6] Deliberate collection of validity evidence (information that provides support for the intended interpretation and use of test scores) helps test developers support their inferred linkages between test scores and the construct being assessed.[2,3,6]

DOI: http://dx.doi.org/10.4300/JGME-D-18-00265.1

How You Can Start TODAY

1. **Create a blueprint.** Determine what content you will test and the percentage of the examination assigned to each kind. Compare the content of existing examinations to your blueprint. Is any content overrepresented or underrepresented?
2. **Analyze how your items perform.** You can do this by estimating item difficulty and item discrimination.
   - Consider the percentage of examinees who correctly answer the item (*P* value). A good examination has a range of *P* values or item difficulties. If a *P* value is greater than .95, the item is too easy; if it is less than .30, the item is too hard.
   - Often your test software will provide the correlation between an examinee's performance on an item and performance on the test as a point-biserial (r_{pb}) correlation. If not, you can do this in a spreadsheet.
   - Create a high-low table by grouping examinees by overall performance (see TABLE).[5] Tabulate the

TABLE
High-Low Table for Multiple-Choice Question Item (*P* Value = .54)

Examinee	Answer (%)				
	A	B	C	D[a]	E
Upper 25%	0	17	11	68	4
Lower 25%	6	31	15	30	18
Total	5	21	13	54	7

[a] D is the correct answer choice.

frequency of answer choices overall and by group. In the example, high performers chose the correct answer, D, more often than low performers (68% versus 30%). If low performers do better on an item than high performers, the item does not discriminate well and should be revised or replaced.

- Assess the distractors (wrong answer choices) that examinees select. Consider replacing distractors that are never chosen. If a distractor is chosen nearly as often as or more frequently than the correct answer, try to determine why and revise the answer choices appropriately.

3. **Review poorly performing items carefully.** Look for extraneous or unclear information in the stem, which can mislead the examinee toward an incorrect answer. Also review the answer choices: Is there only a single correct answer? Are the other response options plausible but clearly incorrect?

What You Can Do LONG TERM

1. **Learn about MCQs.** Download the National Board of Medical Examiner's open-access resource, *Constructing Written Test Questions*.[6]
2. **Keep piloting and analyzing item/examination performance.** As you develop new items, or if you borrow items from review books or previous examinations, pilot them and perform item analysis to ensure they perform as anticipated. The same is true if an assessment is used in a different context (eg, residents instead of medical students).
3. **If your assessment is high-stakes,** whether for assessment or research, consider consulting with a psychometrician or statistician to help measure the reliability of scores.
4. **Consider if what you want to test equals what you are testing (ie, validity).** Accumulate a body of validity evidence to support the intended interpretation and use(s) of your test scores. For example, compare scores from your test with scores from another type of assessment intended to measure the same construct.

Resources

1. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837.
2. American Educational Research Association, American Psychology Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
3. Lane S, Raymond MR, Haladyna TM, et al. *Handbook of Test Development*. 2nd ed. New York, NY: Routledge; 2016.
4. Coderre S, Woloschuk W, Mclaughlin K. Twelve tips for blueprinting. *Med Teach*. 2009;31(4):322–324.
5. Paniagua M, Swygert K, eds. *Constructing Written Test Questions for the Basic and Clinical Sciences*. 4th ed. Philadelphia, PA: National Board of Medical Examiners; 2016.
6. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;19(2):166.e7–e16.

Doug Franzen, MD, MEd, is Associate Residency Director and Assistant Professor of Medicine, Department of Emergency Medicine, University of Washington School of Medicine; **Monica M. Cuddy, MA,** is Measurement Scientist, National Board of Medical Examiners; and **Jonathan S. Ilgen, MD, MCR,** is Associate Professor of Medicine, Department of Emergency Medicine, University of Washington School of Medicine.

Corresponding author: Doug Franzen, MD, MEd, University of Washington, Department of Emergency Medicine, 325 Ninth Avenue, PO Box 359702, Seattle, WA 98104, 206.744.8475, fax 206.744.4097, franzend@uw.edu