Education in Medicine Journal
2012, VOL 4 ISSUE 1
DOI:10.5959/eimj.v4i1.29

EDITORIAL

Question vetting: theory and practice

Majed Wadi

Medical Education Department, School of Medical Sciences, Universiti Sains Malaysia.

How to cite this article?

Wadi, M. (2012). Question vetting: theory and practice. *Education In Medicine Journal*, *4*(1). DOI:10.5959/eimj.v4i1.29

Ensuring and maintaining high quality standards in medical schools have become mandatory tasks (1). This includes one important area, which is assessment of medical students. From this viewpoint, vetting of assessment tools is a crucial process. It is the process of reviewing and evaluating question items according to specified criteria with the intention to detect flaws and to edit them accordingly to improve their quality (2-5). This step is important to sustain the validity of test items and avoid the threats. When we speak about test item quality, we cannot ignore test items validity. Validity is a broad term, though a unitary concept. Sources to support test items validity are several; the character of test items, qualifications of item writers, quality control of securing, scanning, scoring and reporting of tests/exams are all evidences needed to prove content-related validity (6, 7). At the same time, different validity threats should be minimized. Flawed or badly written items are the major threat. They are frequently encountered in many in-house tests (8, 9). However, transforming this theoretical framework to real practice is still not well established.

After evaluation of the quality of in-house examinations used in three American medical schools; in which the majority of them were of relatively low quality, a strong recommendation was made to establish review committees to improve item quality (10). In response to this recommendation, another evaluation was conducted to investigate the effectiveness of the committee review process in improving the quality of in-house examinations. It was found that such a process improved item quality significantly (11).

Most items, even those produced by experienced item writers are still flawed in some ways (5). So, once an item is constructed, it should undergo a critical review by a review (or vetting) committee. The aim is to treat test items appropriately by removing flaws and making them as clear and as understandable as possible (4, 5).

The roles of such a committee were described many years ago (12, 13). Haladyna summarised several activities of an item review process (4). They include item-writing principles review; in which items must be ensured to adhere to identified item-writing guidelines, cognitive demand review; in which the cognitive level of items are assessed, content review; in which the content of each item is matched with what is intended to be measured (testing blueprint), editorial review; where items are checked for any errors in spelling, grammar, and punctuation, sensitivity and fairness review; through which personally, culturally and ethnically offensive terms are removed and substituted by suitable ones, answer key check; where each item is checked for accuracy of the correct answer, answer justification; this is done by listening to examinees' view for their choices during test and accepting their choices if clearly justified And think-aloud; in which a comprehensive discussion is done with test takers to identify relevant information about quality of test items, their contents and cognitive levels being measured.

One of its important tasks is item-writing principles review. There are numerous guidelines of item writing (14-17). One of these is the revised taxonomy of multiple-choice item writing guidelines by Haladyna *et. al.* (16). This taxonomy is based on an extensive review of both educational textbooks and research studies. More than half of these guidelines were supported by experimental studies in item writing. Use of these guidelines during the vetting process is highly recommended. Adherence to these guidelines is considered a source of content-related validity, while violation of any of these guidelines is considered as a construct-irrelevant threat of test validity (18). In two similar studies (8, 19), Downing evaluated the construct-irrelevant variance (CIV) associated with violated (flawed) test items with respect to examination difficulty and pass-fail decisions. He found that violated items were more difficult (higher failure rate) than non-violated items. However, discrepant findings were noticed when such studies were replicated. It was found that non-violated items were associated with lower passing and higher failure rate than violated items (20). In an attempt to standardize the vetting session and make it evidence-based, a study was done to assess the feasibility of using such taxonomy. It was found that these guidelines must be paraphrased and simplified to make them user-friendly (21).

Another important task is the editing process. A number of benefits of the editing process have been delineated (4). First, it shows the cognitive task in a clearer manner. Second, grammatical, spelling, and punctuation errors tend to distract examinees from the main purpose of testing. Third, such errors will be badly reflected on item constructors. Different studies done to assess the effectiveness of such activities have come up with contradictory findings. One study investigated the effect of altering the correct response position of MCQs. It found that item performance was inversely affected by changes in correct option placement (22). On the other hand, other studies found no remarkable differences between edited and unedited items (23, 24).

Review of test item cognitive level is an essential task. The professional life of medical graduates requires them to optimize their cognitive abilities. They identify problems, solve them, interpret findings, think critically, and manage their patients comprehensively. Assessment tools should measure up to this level rather than just assessing recall and factual knowledge (25-27). In this regard, the role of the reviewing committee is to assess the items' cognitive level rather than to change or reconstruct them again. In order to improve the cognitive levels of test item, the item developer should reconstruct and re-write the test items(4). It was found that majority of violated items test recall and factual knowledge (28).

From the psychometric point of view, test item review improved item difficulty and discrimination indices (29). In one study (30), the effect of reviewing and improving discarded items on the item discrimination index was evaluated. Considerable

DOI:10.5959/eimj.v4i1.29

improvement of the item discrimination index was found after the reviewing process. Downing found that flawed items (violated) were more difficult and had lower discriminative ability than unflawed (non-violated) items (19, 31).

Based on the previous highlighted points, it can be said that question vetting is a major part of quality insurance in medical education. However, it consumes a substantial amount of time and effort. More research to come up with a standardised vetting process and make it more achievable with less time and resources is highly suggested. Evidence-based practice is crucial to both administrative and academic staff. Justified by evidence, the first can confidently make decisions to consume resources in the vetting process, while the latter can evaluate their work and see the merit of such daily practices.

Reference

1. Corrigan O, Ellis K, Bleakley A, Brice J. Quality in m edical e ducation. In: Swanwick T, editor. Understanding Medical Education. 1st ed: Wiley-Blackwell; 2010. p. 379 - 291.
2. American Educational Research Association, American Psychological Association, National Councilon Measurementin Education. Standards for educational and psychological testing. Washington DC: American Educational Research Association.; 1999.
3. Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van Der Vleuten CPM. Quality Assurance in Test Construction: The Approach of a Multidisciplinary Central Test Committee. Education for Health: Change in Learning & Practice (Taylor & Francis Ltd). 1999;12(1):49.
4. Haladyna TM. Developing and validating multiple-choice test items. 3rd ed: Lawrence Erlbaum; 2004.
5. Baranowski RA. Item Editing and Editorial Review. In: Downing SM, Haladyna TM, editors. Handbook of Test Development. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers; 2006. p. 349-57.
6. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003;37(9):830-7.
7. Downing SM, Haladyna TM. Validity and Its Threats. In: Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. 1st ed. New York: Routledge; 2009.
8. Downing S. Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. Advances in Health Sciences Education. 2002;7(3):235-41.
9. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Medical Education. 2004;38(3):327-33.
10. Jozefowicz RFMD, Koeppen BMMDP, Case SP, Galbraith RMD, Swanson DP, Glew RHP. The Quality of In-house Medical School Examinations. Academic Medicine. 2002;77(2):156-61.
11. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. Advances In Health Sciences Education: Theory And Practice. 2006;11(1):61-8.
12. Hubbard JP. Measuring Medical Education Philadelphia: Lea and Febiger; 1971.
13. Anderson J. The Multiple Choice Question in Medicine 2nd ed. London: PITMAN BOOKS LIMITED; 1982.
14. Case S, Swanson D. Constructing Written Test Questions for Basic and Clinical Sciences. 3 ed. Philadelphia National Board of Medical Examiners; 1998.
15. Osterlind SJ. Constructing Test Items: Multiple-Choice, Constructed Response, Performance and Other Formats. 2 ed. New York: Kluwer Academic Publishers; 2002.
16. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education. 2002;15(3):309 - 33.

17. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse Educator. 2012, In Press;37(3).

18. Downing SM. Written Tests: Constructed-Response and Selected-Response Formats. In: Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. New York: Routledge; 2009.

19. Downing SM. The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. Advances in Health Sciences Education. 2005;10(2):133-43.

20. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education. 2008;42(2):198-206.

21. Wadi M, Yusoff MSB, Rahim AFA. The 31 Revised Taxonomy of Multiple-Choice Item Writing Guidelines: Can it be used as a Vetting Protocol? Paper presented at the 15th Ottawa Conference, Kuala Lumpur.2012.

22. Cizek GJ. The effect of altering the position of options in a multiple-choice examination. Educational and Psychological Measurement. 1994;54(1):8-20.

23. O'Neill KA. The Effect of Stylistic Changes on Item Performance. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.1986.

24. Webb L, Heck W. The effect of stylistic editing on item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago1991.

25. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. Academic Medicine. 1981;56(2):115-21.

26. Ferland JJ, Dorval J, Levasseur L. Measuring higher cognitive levels by multiple choice questions: a myth? Medical Education. 1987;21(2):109-13.

27. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Academic Medicine. 1990;65(9):S59-60.

28. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education Today. 2006;26(8):662-71.

29. Ebel RL, Frisbie DA. Essentials of Educational Measurement. 5th ed: Prentice-Hall, Inc., Englewood Cliffs, New Jersey; 1991.

30. Lange A, Lehmann IJ, Mehrens WA. USING ITEM ANALYSIS TO IMPROVE TESTS. Journal of Educational Measurement. 1967;4(2):65-8.

31. Downing SM. Construct-irrelevant Variance and Flawed Test Questions: Do Multiple-choice Item-writing Principles Make Any Difference? Academic Medicine. 2002;77(10)(Supplement):S103-S4.

Corresponding Author

Dr Majed Mohammed Saleh Wadi
Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia.
16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
Email: majed_wadi@yahoo.com