Does central vetting improve construct quality of one-best-answer items in medical school: An audit

Siew Kim Kwa[1], Zainab Majeed[2], Shane Varman[2]

Introduction: Assessment is an integral aspect of teaching. One-best-answer (OBA) items, if properly constructed are able to drive learning. In-house OBA items are notoriously poorly-constructed. The role of a central vetting committee is to review test items and ensure that they adhere to expected standards. Hence, the objective of this audit is to determine whether central vetting has improved the construct quality of OBA items.

Methods: We audited the psychiatry end-of posting OBA items from before and after central vetting to compare the quality of the items before and after central vetting was instituted. Quality was evaluated on appropriateness of test content, items with higher cognition and items without flaws. A standard was not set for this first audit.

Results: Seventy six of 181 psychiatry OBAs items retrieved from 2011 to August 2012 had undergone first level (department) vetting only and the remainder 105 (58.0%) had two levels of vetting; department and central vetting committee (CVC).

Appropriateness of content increased from 92.1% to 98.1%. Items with higher order thinking doubled from 21.1% to 42.9%. Items with clinical scenario increased by 8.4% to 78.1%. Logical ordering of options however, remained around 50%.

Two-level vetting markedly reduced problematic lead-in questions (67.1 to 13.3%), non-homogenous options (42.1 to 9.5%), vague and implausible options (39.5 to 6.7%), and spelling and grammar mistakes (19.7 to 5.7%).

Conclusion: Two-level vetting had improved the quality of OBAs and should be continued. This could be enhanced by training all Faculty on writing quality OBA items and careful selection and empowerment of CVC members. A re-audit is to be conducted after Faculty training.

IeJSME 2017 11(3): 3-9

Key words: Assessment, Vetting, One-best-answer items, MCQ, Quality assurance

Introduction

Assessment is an integral aspect of teaching. It can be formative, in-course for learning or summative, to determine attainment of competency for the next level.[1,2] Formative assessment helps guide students on the learning outcome expected. In many medical schools, multiple choice questions (MCQs) are fast replacing essay questions because they can test across a broad range of topics in the curriculum.[3] MCQs are cost-efficient as there is no limit to the number and location of candidates and can be conducted with minimal or no supervisory staff. They can be promptly computer-scored without examiner bias, thus facilitating timely feedback to candidates.

Nevertheless, MCQs have come under much criticism. The True/False MCQs used previously tend to test rare and obscure facts at low cognitive levels of knowledge and understanding[3] whereas medical students need to develop higher cognitive levels of clinical reasoning, application and decision-making. Hence, the rise of the ubiquitous one-best-answer (OBA) or commonly called single-best-answer item. This consist of a stem, lead-in question and 4 – 5 response options of which one, is clearly most correct.

Assessment drives learning.[1,2] The future medical students must be able to manage patients. Therefore, assessment should mirror real-life problems faced daily by doctors and include relevant clinical scenarios. The lead-in question should focus on clinical features, pathophysiology, investigations, diagnosis or management. Response options should preferably be short and not contain cues, be imprecise or have absolute counts. The correct answer should clearly be the best choice and there should be no grammatical or spelling mistakes. Various papers have spelt out rules and guidelines for constructing quality OBA test items.[3-7]

[1]Department of Family Medicine, International Medical University, 70300 Seremban, Negeri Sembilan, MALAYSIA
[2]Department of Psychiatry, International Medical University, 70300 Seremban, Negeri Sembilan, MALAYSIA

Address for Correspondence:

Prof Dr Kwa Siew Kim, Department of Family Medicine, International Medical University, Jalan Rasah, 70300 Seremban, Negeri Sembilan, MALAYSIA
Email: siewkim_kwa@imu.edu.my

Non-adherence to the accepted rules is considered an "item-writing flaw" (IWF).

OBA items, if well written, can test higher cognitive functions of clinical reasoning, application of knowledge and decision-making skills expected of the future doctors. But OBA items constructed in-house for low-stake exams have been found to be less than satisfactory and contain many IWF.[8,9] Poorly constructed OBA items can adversely affect students' examination performance.[10] Items without a clinical scenario will not drive student's learning towards the outcome required of the future doctor.

As part of quality assurance (QA) for assessments, all test items must be vetted for construct quality.[11] Good construct quality in test items does not mean only absence of IWF but should also include quality indicators of higher cognition, appropriateness, relevance and adherence to curriculum and assessment blueprint.

At the International Medical University (IMU) in Malaysia, individual Departments in the School of Medicine (Clinical campus), generate OBA items for the end-of-posting (EOP) examinations. Vetting at the first level is conducted at the department. A second level of vetting centrally at University level was introduced for all examinations from August 2012. The role of the central vetting committee is to ensure that all OBA items are without writing flaws and meet guidelines for the construction of good OBA items.[3-7]

The objective of this paper is to audit and evaluate whether the additional central vetting at University level has improved the quality of the OBA items used in the EOP examination.

Methods

We selected all the Psychiatry OBA items used in the EOP examinations for Semester 7 (Year 3) from 2011 to 2015 in the IMU. Earlier items were only vetted at Department level but from August 2012 to 2015, all items underwent an additional vetting by a Central Vetting Committee (CVC) appointed by the Dean's office.

We evaluated the quality of the OBA test items on three broad aspects: appropriateness of test item content for students' level of learning, cognitive level and IWF.

Appropriateness of Test Item

The topic and content chosen for the OBA item should be from within the assessment blueprint for the curriculum and the subject matter is relevant and/or important for the learners at that posting.

Cognitive level

As practised by Ware,[11] we created a 2-tier system (K1 and K2) from the original 6-level Bloom's taxonomy on cognition (Figure 1). We collapsed the lowest two levels 'remembering' and 'understanding' under Tier 1 for 'recall of knowledge' (K1). The other four higher cognitive levels of 'application, analysis, evaluation and creation' were aggregated under 'clinical application' (K2).

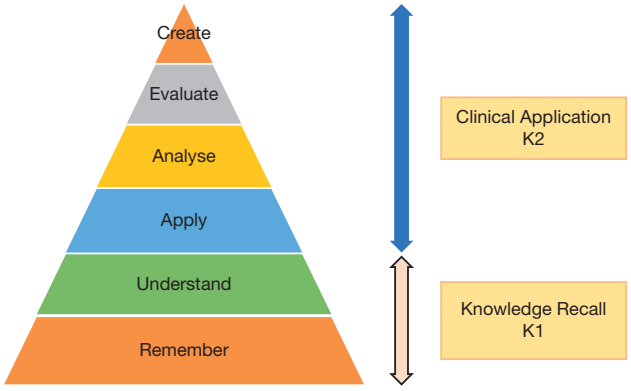

Figure 1: Revised Bloom's Cognitive levels: K1 and K2

Item Writing Flaws (IWF)

A check list for IWF was created based on the recommended principles for writing good OBA items.[3-7] Table I summarises the anatomy of the ideal OBA items. This should consist of a stem with a clinical vignette detailing relevant information, a clear succinct lead-in question and four to five short response options. A clinical scenario is essential for all OBA items at clinical school level to achieve the final objective of training medical students into competent doctors capable of managing clinical problems.

Table I: Recommended Guidelines for a well-constructed One-Best-Answer Item

Stem (Vignette) An 18 year old girl has fear of crowded areas and refuses to leave home.	Contains relevant clinical data to answer the lead-in question
Lead-in Question What is the most appropriate non-pharmacological management?	Must be succinct and can be clearly understood by students Should pass the cover test
Options A. Cognitive behaviour therapy B. Desensitization C. Interpersonal therapy D. Psychoanalysis E. Supportive therapy *A is the correct answer* *B to E are distractors*	**Options** – Usually short – Homogenous – Plausible distractors – Ordered logically – No vague terms (sometimes, maybe) – No absolute terms (all, none) – No clues /cues

**There should be no spelling or grammatical errors in the item*

The five options consist of one single best answer and four "distractors". Options should be short, homogenous e.g. diagnoses, investigations, etc. with approximately similar lengths. Distractors should be plausible but may include common misconception. Ideally the item can be answered without looking at the options (i.e. pass the cover test). There must be no obvious clues or cues. Options should not be absolute e.g. "all of the above", "none of the above. They should not be imprecise or vague e.g. may, often, sometimes, usually, etc. The options should be placed in alphabetical or numeric order for ease of answering and to prevent cueing. There should be no spelling or grammatical mistakes and non-universal abbreviations.

IMU's research and ethical committee recommended a minimum of at least three research team members, two of whom should be Psychiatry content experts to evaluate the content of the items for appropriateness to the curriculum and students' level of learning. Each research team member evaluated the quality of the OBA items independently but met to discuss and decide on a final consensus.

Results

A total of 181 OBA items were evaluated; 42.0% (76 items from 2011 to February 2012) had undergone only first level vetting at the department level and 58.0% (105 items from August 2012 till 2015) received a second vetting at the central university level (Table 2).

Table 2: Changes in Quality Indicators after Central Vetting

Items	Before Central Vetting *N1=76	After Central Vetting Instituted **N2=105
Items with clinical scenario	53 (69.7%)	82 (78.1%)
Items with appropriate content	70 (92.1%)	103 (98.1%)
Items test higher cognition	16 (21.1%)	45 (42.9%)
Item-writing Flaws		
Lead-in question not clear	51 (67.1%)	14 (13.3%)
No logical ordering of options	38 (50.0%)	53 (50.5%)
Implausible and vague options	30 (39.5%)	7 (6.7%)
Non-homogenous options	32 (42.1%)	10 (9.5%)
Spelling and grammar mistakes	15 (19.7%)	6 (5.7%)

**N1 is the total number of items without central vetting*
*** N2 is the total number of items that had undergone central vetting*

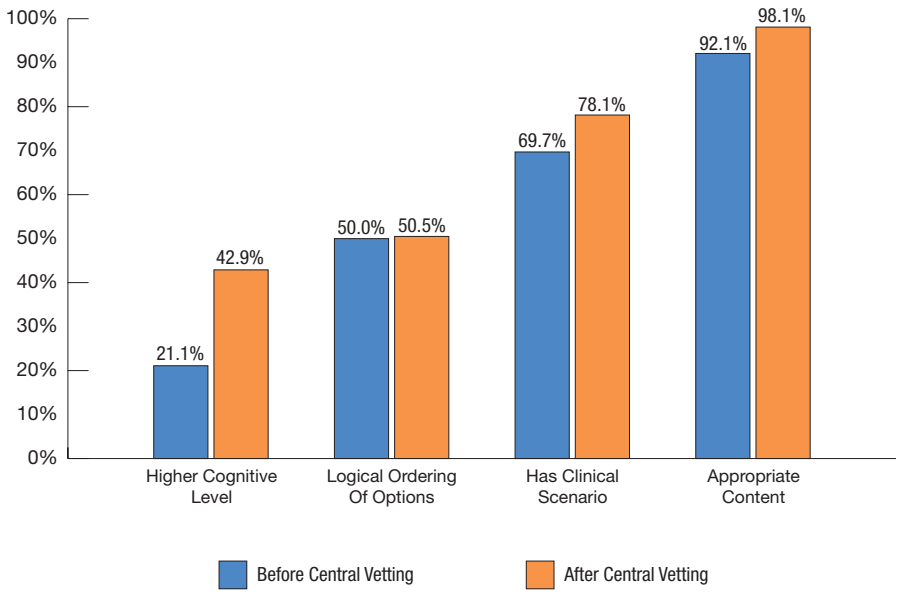

Figure 2: Improvement in Quality Indicators after Central Vetting

Figure 2 shows an impressive doubling to 42.9% of items with higher order cognitive levels after central vetting. However, there was minimum change in logical ordering of options at about 50%. Items with clinical scenario increased only by 8.4% to 78.1% except in the February 2015 examination, when every item had a clinical stem. Appropriateness of item content to students' assessment, already very high at 92% increased further after central vetting but did not reach the expected 100%.

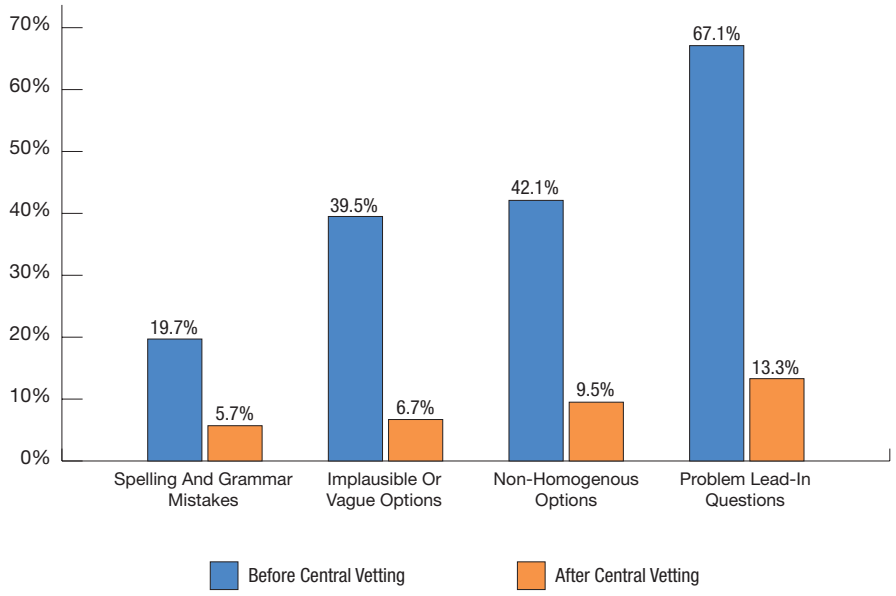

Figure 3: Reduction in Item Writing Flaws (IWF) after Central Vetting

Figure 3 demonstrates a marked decrease in IWF after two-level vetting was instituted; problematic lead-in questions (67.1 to 13.3%), non-homogenous options (42.1 to 9.5%), implausible or vague options (39.5 to 6.7%), and spelling and grammar mistakes (19.7 to 5.7%).

Discussion

Pre-test vetting by department Faculty members and re-vetting by the CVC are important processes to meet accreditation requirements for quality assurance in assessments.[11,12,13] Credit should be given to the Psychiatry department for producing more than 90% of OBA items with appropriate learning content even without central vetting.

This paper demonstrates that a second vetting at the central level plays an important role in reducing technical and language flaws overlooked by the department. Although items with higher cognitive order of clinical application had doubled to 43% with central vetting, this is not good enough as ideally, it should be 60-80% for a clinical school.

Logical ordering of response options remain unchanged at 50% despite a second vetting. A possible explanation is that the CVC are not aware of this rule

or do not consider it important enough to warrant their attention. But logical ordering of options is very important to prevent cueing for the exam-savvy candidates. One of the hallmarks in MCQ examination is that of speed in arriving at the correct answer. Logical ordering helps students to quickly identify the required answer from the list of response options.

We would have expected that the CVC would have identified most flawed items and rejected all items without a clinical scenario to test clinical reasoning and decision-making. One possible explanation for the continual presence of IWF despite two-level vetting could be due to late submission of additional test items to replace flawed ones. Due to tight examination timelines, central vetting could have been bypassed.

In February 2015, the CVC ensured that every OBA item had a clinical scenario. This means that adherence to guidelines for quality construct of test items is possible but require selection of CVC members who are well-versed with the guidelines and enforcement. The quality of vetting depends substantially on the capacity and capability of its members. Both Shahid[12] and Gopalakrishnan[13] recommended a formal structure for the vetting process. CVC members who are well-versed with the principles of good test item construct should be formally appointed by the University and the Chairperson and empowered to enforce the regulations. CVC members should be well-trained in evaluating and detecting flawed test items and at least one content expert must be present during the discussion.

Due to teaching and other administrative duties, CVC members may not all be present at every vetting session. It is recommended that to maintain quality assurance in assessments and to reduce vetting time, all Faculty should be trained and given a chance to be part of the CVC to enable them to be familiar with recommended guidelines and the vetting procedure. Otherwise faced with time constraints, CVC may just be confined to evaluating language and ignoring technical and content issues.

In conclusion, second-level central vetting is essential and does make a significant difference in constructing quality test items for quality assurance in assessment. The use of a prescribed structure and protocol of question setting, deadline for submission and vetting can ensure a consistently high standard of assessment items. It is recommended that for continuing quality assurance in assessment, all Faculty should be trained on writing quality items. The Faculty involved will be informed of the audit results and the recommendations for vetting and training. Following a time span of about a year, another audit will be conducted to look for further improvement.

Acknowledgements

The authors wish to acknowledge the assistance of the IMU Academic Services Department in providing the examination papers. We also thank IMU Joint Research and Ethics Committee for their approval and funding to conduct this study (approval number IMU 368/2016).

REFERENCES

1. Ferris H, O'Flynn. Assessment in Medical Education; What are we trying to achieve? Int Jour Higher Educ 2015; 4:139-44.
2. O'Farrell C. Enhancing Student Learning through Assessment. Dublin Institute of Technology. 2005. Retrieved from http://www.tcd.ie/teaching-learning/academic-development/assets/pdf/250309_assessment_toolkit.pdf [Accessed on 4 Aug 2016]
3. Case S, Swanson D, eds. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd edn. National Boards of Medical Examiners, Philadelphia. 2002. http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf [Accessed: 4 July 2016]
4. Haladyna TM, Downing, SM. A taxonomy of multiple choice item-writing rules. Applied Measurement in Education 1989; 2(1):37–50.
5. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 1989; 2(1): 51–78.
6. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 2002; 15 (3), 309–34.
7. Wood T, Cole G, eds. Developing Multiple Choice Questions for the RCPSC Certification Examinations. Educational Research and Development, Royal College of Physicians and Surgeons of Canada, June 2004. http://www.ranzcog.edu.au/fellows/pdfs/diploma-mcqs/developing-mcqs-for-RCPSC.pdf
8. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med 2002; 77:156-61.

9. 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 in Practice 2006; 6: 354-63.
10. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed items on achievement examinations in medical education. Adv Health Sci Educ 2005; 10: 133-43.
11. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Medical Teacher 2009; 31:3
12. Shahid H, Nordin S, Harmy MY. Structured Vetting Procedure of Examination Questions In Medical Education In Faculty Of Medicine at Universiti Sultan Zainal Abidin Malaysia. Malaysian Journal of Public Health Medicine 2016; 16(3): 29-37.
13. Gopalakrishnan S, Udayshankar PM. Question Vetting: The Process to Ensure Quality in Assessment of Medical Students. Journal of Clinical and Diagnostic Research 2014; 8(9): XM01-XM03. DOI: 10.7860/JCDR/2014/9914.4793