Assessment & Evaluation in Higher Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caeh20

Peer review improves the quality of MCQ examinations
Bunmi S. Malau-Aduli a & Craig Zimitat a
a School of Medicine, University of Tasmania, Hobart, Australia

To cite this article: Bunmi S. Malau-Aduli & Craig Zimitat (2012): Peer review improves the quality of MCQ examinations, Assessment & Evaluation in Higher Education, 37:8, 919-931
To link to this article: http://dx.doi.org/10.1080/02602938.2011.586991

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Peer review improves the quality of MCQ examinations

Bunmi S. Malau-Aduli* and Craig Zimitat

School of Medicine, University of Tasmania, Hobart, Australia

The aim of this study was to assess the effect of the introduction of peer review processes on the quality of multiple-choice examinations in the first three years of an Australian medical course. The impact of the peer review process and overall quality assurance (QA) processes were evaluated by comparing the examination data generated in earlier years (2008) with those held under the new QA regime (2009 and 2010) from the same blueprint. Statistical analysis and comparisons of overall examination performance were made by year. Regarding multiple-choice questions (MCQs), item analysis was used to compare the proportion of difficult and discriminating items and functional distractors on summative examinations in 2008 (pre-implementation of peer review) and 2009 and 2010 (post-implementation). The impact of peer review processes resulted in a decrease in the number of items with negative discrimination; increases in reliability, appropriate item difficulty, and numbers of items with significant discrimination. There was an associated improvement in the effectiveness of distractors for the MCQ items. The trend of overall improvement in the quality of MCQ items continued in 2009 and 2010. The introduction of QA processes, specifically peer review of MCQ items has resulted in a sustained improvement in the quality of MCQ items within our examinations.

Keywords: assessment; quality assurance; multiple-choice questions; item analysis; test construction

Introduction

Evaluating the quality of any educational enterprise requires evaluation of the quality of the assessment within that system. Rowntree (1987) and Ramsden (2003) highlight the pervasive positive and negative influences that assessment strategies have on learning, particularly if they are misaligned with educational philosophies and desired learning outcomes of a course. Maximising the positive educational impacts of assessment requires attention to the overall programme philosophy and goals, balancing of practicalities of assessment methods with reliability and validity (van der Vleuten 1996) and using processes and data to determine the success of the interventions; and communicate the results of the assessment process (McLachlan 2006) ultimately to improve learning.

Evaluating the quality of assessment within a course requires review at multiple levels. At the institutional level, policies, guidelines and procedures set the overall framework for assuring integrity of the assessment system, and peer review and scrutiny of assessment choices as they relate to educational outcomes (Baartman, Bastiaens, and Kirschner 2006). At the school and course levels, processes such as

*Corresponding author. Email: bunmi.malauaduli@utas.edu.au
blueprinting, balancing formative and summative assessment strategies, choice of instruments, staff training and feedback processes also affect student performance and learning. At the individual level, personal choices related to the selection of subject matter, engagement in training and review and overall commitment affect the quality of assessment items offered within the system.

A key matter to address in the design of an assessment system is the attention to validity and reliability. Validity is a global construct related to the inferences that can be drawn from the outcomes of an assessment, through examination of multiple sources of data (Downing and Yudkowsky 2009). It includes qualitative ‘external’ processes as much as blueprinting (Hamdy 2006), training etc. that relate to the environment in which assessment is conducted. Reliability is a quantitative indicator, intrinsic to the test that defines the reproducibility of assessment over time (Downing and Yudkowsky 2009). The two concepts are ‘traded off’ in the design of assessment to ensure that assessments measure reliably (enough) what is intended (or actually) to be measured: reliability is a necessary but insufficient pre-condition for validity (Schuwirth and van der Vleuten 2006). Overall, without validity, few conclusions can be drawn from an assessment to adequately inform decision-making based upon that assessment data.

Multiple choice questions (MCQs) are widely used assessment instruments. They are considered as easy to write, capable of assessing higher-order thinking and efficient when used with large classes and across multiple campuses. MCQ tests, if long enough, can assess the full range of content areas with high reliability (Epstein 2007). However, there is considerable evidence that MCQs are poorly written and too frequently assess recall rather than higher-order thinking (Vahalia et al. 1995) because of inadequate attention to the writing process (Josefowicz et al. 2002). Large numbers of internet and printed resources are available to support the writing of quality MCQs (Case and Swanson 2003; Downing and Haladyna 2006). Poor quality items can have detrimental effects on learning and the reliability and validity of assessment (Downing 2005). The validity of MCQ tests and the quality of MCQ items can be improved by the use of templates, qualitative review processes and the review of questions based upon statistical analysis of assessment data (Downing and Haladyna 1997). National (e.g. National Board of Medical Examiners – NBME) and, more recently, international consortia (such as International Database for Enhancement of Assessment and Learning – IDEAL and Universities Medical Assessment Partnership – UMAP) have developed databases and processes to provide members with access to quality assessment items (e.g. Best of ‘N’ options MCQ, Short Answer Questions – SAQ, Extended Matching Questions – EMQ and Objective Structured Clinical Examination – OSCE) as broad strategies to address assessment quality in medical education.

Meaningfully interpreted assessment scores require content-related validity evidence of the adequacy of the content tested and statistical evidence of score reproducibility and item statistical quality (Downing 2004). Item analysis is a quality control tool for tests, providing quantitative data at the item-level, as well as some important summary statistics about the whole test. Careful review of item analysis data can help to improve the reliability and consequently the validity of scores generated by instruments. Item analysis using classical test theory defines a number of quantitative characteristics of quality MCQ items. At least four different analyses are relevant for evaluating the quality of MCQs (Downing 2003). Item difficulty (DIF), also known as $p$-value, refers to the proportion of examinees who answered the
question correctly, with lower values reflecting potentially more difficult questions (Osterlind 1998). DIF is a measure of validity as higher values indicate the proportion of students who have learned the content measured by the item. The discrimination index (DI) is an indicator of how items are answered by high and low-achieving students within a test. DI is a basic measure of the validity of an item. It can be interpreted as an indication of the extent to which the overall knowledge of the content area or mastery of the skills is related to the response of an item. Taking the DI one step further, the point-biserial correlation (RPB) provides a correlation between the score that students receive on a given item (right/wrong) and the total score for the examination. This reflects the fact that reliability involves freedom from random error, thus, the less the random error in the variables, the higher the possible correlation between them (Downing 2002). Researchers (Downing and Haladyna 2006; McAlpine 2002) have made suggestions on recommended values for item difficulty and discrimination based on the theory that the most informative test items are those of middle difficulty and they provide higher discrimination between the high-scoring and low-scoring examinees. It is noted that adjustments ought to be made to these recommendations for educational courses based on a competency or mastery philosophy (Downing and Yudkowsky 2009). Distractor analysis distinguishes between functional distractors (i.e. plausible distractors chosen by examinees) and non-functional distractors that are rarely selected by candidates. It provides an indication of the quality of the question, in particular the options that relate to the question posed in the stem. The fewer the functional distractors the greater the chance examinees have of selecting the right answer regardless of their knowledge of the topic. The greater the number of plausible distractors, the more accurate, valid and reliable the test becomes (Haladyna and Downing 1993).

There are strong arguments and recommendations for collaborative review processes in test development (Downing and Haladyna 2006) as well as many resources and materials available to support examiners to write quality MCQ. However, there is little empirical evidence that relates the use of educational and training resources and processes to the improvement of the quality of MCQs. The NBME (Case and Swanson 2003) have produced an extensive guide for writing MCQs for the basic sciences that provides guidance on consistency of formats, word usage, structure and avoidance of common writing flaws. Editing to improve English language and terminology leads to small but significant effects on item performance (Dawson-Saunders et al. 1992). Josefowicz et al. (2002) studied assessment in three US medical schools and reported that training assessors in writing MCQs significantly increased the quality of items when judged by experts from the National Board of Medical Examiners. A similar US study by Wallach et al. (2006) demonstrated that the use of templates, style guides and peer review improved the quality of MCQ items. In each of these studies, the quality of MCQ items after peer review was assessed by expert panels; the effect on the quality of items assessed after use in an examination have not been investigated. This paper reports the effect of the introduction of peer review processes on the quality of multiple-choice examinations as determined by statistical analysis.

**Organisational context of this study**

The University of Tasmania medical course is a five-year undergraduate, integrated case-based programme. Vertical integration is supported by the use of a thematic
structure across the five years of the programme. The first two years of the course provide a systems-based introduction to the foundations of medicine, with early experiences in the development of communication and clinical skills. In Year 3 of the programme, students commence clinical rotations in medicine, surgery, primary care and specialty areas (obstetrics and gynaecology, psychiatry and paediatrics), whilst the final two years of the programme comprise hospital-based rotations. Assessment in the first three years is blueprinted, and uses a range of instruments including MCQ, SAQ, EMQ, OSCE and Portfolio. MCQ examinations in this study were held in three semester-long modules: MED101 – Foundations of Medicine (13 weeks), MED201 – Foundations of Cardiovascular and Respiratory medicine (13 weeks) and MED301 – Foundations of Neurology and Psychiatry (18 weeks).

The Medical Education Unit (MEU) is central to the management of the assessment process for the MBBS programme. In 2008, the MEU commenced a process of rationalisation and systematisation of assessment processes that ‘wrapped around’ the IDEAL assessment database (Figure 1). The IDEAL assessment database is a Windows-based software application for entering, storing, retrieving, editing, printing or administering assessment items as well as for updating, managing and sharing assessment banks in medical education (IDEAL 4.1-HK 2003). It maintains a formative and a summative bank of assessment items which span all years, disciplines and systems normally addressed in an initial medical degree programme. The Tasmanian School of Medicine (TSoM) is one of the 27 international members of the IDEAL Consortium, which is a group of medical schools committed to enhancing the assessment of medical students. The model encompassed the management of exam writing, exam preparation and management, statistical analysis of examina-

Figure 1. Overview of the relationships between the IDEAL database and QA processes.
tion data and introduction of automated feedback processes to staff on examination items and to students about their examination performance. Quality assurance processes (Malau-Aduli, Zimitat, and Malau-Aduli 2011) were developed in draft format by the MEU, followed by a process of individual consultation, discussion, review and ratification at the annual Curriculum Forum with minor modifications for implementation in 2009. The QA processes around assessment practices included blueprinting the educational objectives, selecting appropriate test formats and applying assessment strategies to achieve adequate levels of reliability. It also included implementing appropriate standard-setting and decision-making procedures and running workshops for the peer review of assessment items before they were administered to students to ensure that the items are constructed properly, without item-writing flaws. To minimise the error of measurement and thus increase the overall validity and reliability of the exam, the review workshops focused on: (1) what the question tested, (2) why the question was relevant or important, and (3) if it was core knowledge.

Methodology

Sample description
All of the summative MCQs administered to Year 1–3 undergraduate medical students at the TSoM (in MED101, MED201 and MED301) for the academic years 2008, 2009 and 2010 were retrieved from the database for analysis. These MCQ were developed by academic and clinical staff within the medical school, as well as from staff in consortium member schools. All items were five-option, single best answer questions with no penalty for incorrect answers. The 2009 and 2010 MCQ items were blueprinted and reviewed by colleagues and assessment committees before use (Figure 1), but the 2008 MCQ items were not peer-reviewed during the writing process, or after the examination review. With the introduction of the peer review process, item writers were encouraged to develop items that assessed high cognitive-level processes rather than recall of factual knowledge. All assessment items were classified using the IDEAL classification scheme (IDEAL 4.1-HK 2003) and stored in the item bank.

Psychometric measurements and statistical analysis
Item analysis was conducted for each assessment by computing the difficulty index (DIF), DI, item-total score correlation coefficients (RPB) and Kuder–Richardson 20 (KR-20) reliability using the classical test theory as provided in IDEAL 4.1, an Item Analysis Program (Precht et al. 2003). The quality criteria chosen for each of these quality indicators were derived from McAlpine (2002). The target range for DIF was defined by \( p \)-values \( \geq 20 \) to \( \leq 80 \). Items with discrimination indices ranging from 0 to 0.15 were defined as non-discriminating, while those with values ranging from 0.16 to 1.0 were classified as discriminating. Non-functioning options were defined as those chosen by fewer than 5% of examinees (Haladyna and Downing 1993; Osterlind 1998; Tarrant and Ware 2008; Tarrant, Ware, and Mohammed 2009). All distractors with a choice frequency <5% and positive discriminating power were identified. Frequency distributions were constructed for items with 0, 1, 2, 3 and 4 functioning distractors.
A general linear models (GLM) procedure in SAS (SAS 2002) was utilized in computing least square means and their associated standard errors (SE) and descriptive statistics of the variables. Means of item difficulty, DI and number of functioning distractors per item in all exams were computed using the least significant difference technique.

Results

Least square means estimates on psychometric parameters are presented in Table 1. There were significant differences ($p < 0.001$) in item difficulty (DIF), DI, point-biserial correlation (RBP) and functional distractors (FD) in relation to year of exam, with all 2008 tests having higher DIF and lower DI, RPB and FD than the 2009 and 2010 tests.

Quality indicators were determined for each of the MCQ examinations in each year of this study (Table 2). The total number of MCQ items was 866 (range: 65–130 items per test), while the total number of examinees was 989 (range: 55–125 examinees per test). Mean test scores ranged from 62% to 73% with higher mean scores in 2008 compared to 2009 and 2010 (range: 64–73%, 60–68% and 62–67%, respectively). The percentage of items considered too easy (DIF $\geq 80$) or too difficult (DIF $\leq 20$) decreased in all tests held in 2009 and 2010 compared to those held in 2008. The mean DI ranged from 0.17 to 0.25 in 2008, increasing in range from 0.26 to 0.29 in 2009 and from 0.24 to 0.27 in 2010. Internal consistency reliability of test scores ranged from 0.64 to 0.81 in 2009 and 0.72–0.80 in 2010. The percentage of items that assessed recall of factual knowledge (based on classification) decreased from 65% in 2008 to 31% in 2009 and 30% in 2010. The percentage of items with negative or low DI decreased in all tests held in 2009 and 2010 compared to those held in 2008.

Each MCQ in the test comprised a stem (question) and five options including the key (correct answer) and four distractors (Table 3). Distractor analyses were completed for the test items to identify non-functional distractors. Overall, 866 items with 3464 distractors were assessed. Of the total number of distractors, 450 (13.1%) were not chosen by any of the examinees. Across the three tests, 21% of the distractors were not chosen by any examinee (i.e. the answer key was obvious) in 2008, compared to 9% in 2009 and 10% in 2010. Conversely, the percentage of functional distractors overall increased from 44% in 2008 to 57% and 54% in 2009 and 2010, respectively. The overall mean number of functioning distractors per item was 2.07, with ranges of 1.32–2.15 in 2008 compared to ranges of 2.02–2.41 in 2009 and 2.00–2.32 in 2010. There was an increase in the percentage of items with
<table>
<thead>
<tr>
<th>Criteria (%)</th>
<th>MED101</th>
<th>MED201</th>
<th>MED301</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items in exam</td>
<td>110</td>
<td>130</td>
<td>127</td>
</tr>
<tr>
<td>Number of examinees</td>
<td>119</td>
<td>123</td>
<td>124</td>
</tr>
<tr>
<td>Mean score % (SD)</td>
<td>64 (7.83)</td>
<td>60 (6.99)</td>
<td>62 (7.54)</td>
</tr>
<tr>
<td>Mean difficulty % (SD)</td>
<td>73.93 (24.51)</td>
<td>67.57 (19.32)</td>
<td>68.14 (22.57)</td>
</tr>
<tr>
<td>Mean discrimination index (SD)</td>
<td>0.17 (0.15)</td>
<td>0.26 (0.15)</td>
<td>0.25 (0.15)</td>
</tr>
<tr>
<td>Reliability index</td>
<td>0.61</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Easy items (%)</td>
<td>61 (55)</td>
<td>36 (28)</td>
<td>46 (36)</td>
</tr>
<tr>
<td>Difficult items (%)</td>
<td>7 (6)</td>
<td>1 (1)</td>
<td>6 (5)</td>
</tr>
<tr>
<td>Recall items (%)</td>
<td>68</td>
<td>33</td>
<td>39</td>
</tr>
<tr>
<td>Items with negative discrimination indices (%)</td>
<td>10 (9)</td>
<td>5 (4)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Items with zero discrimination indices (%)</td>
<td>6 (5)</td>
<td>2 (2)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Items with low discrimination indices (%)</td>
<td>31 (28)</td>
<td>22 (17)</td>
<td>31 (24)</td>
</tr>
<tr>
<td>Total no. of discriminating items (%)</td>
<td>63 (57)</td>
<td>101 (78)</td>
<td>90 (71)</td>
</tr>
</tbody>
</table>
Table 3. Distractor analysis.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items in exam</td>
<td>110</td>
<td>130</td>
<td>127</td>
<td>65</td>
<td>74</td>
<td>114</td>
<td>67</td>
<td>84</td>
<td>95</td>
<td>866</td>
</tr>
<tr>
<td>No. of distractors assessed</td>
<td>440</td>
<td>520</td>
<td>508</td>
<td>260</td>
<td>296</td>
<td>456</td>
<td>268</td>
<td>336</td>
<td>380</td>
<td>3464</td>
</tr>
<tr>
<td>Distractors with frequency = 0% n (%)</td>
<td>96 (21.8)</td>
<td>71 (13.7)</td>
<td>65 (12.8)</td>
<td>33 (17.5)</td>
<td>17 (9.2)</td>
<td>42 (7.4)</td>
<td>74 (27.6)</td>
<td>24 (7.1)</td>
<td>28 (13.1)</td>
<td>450</td>
</tr>
<tr>
<td>Distractors with frequency &lt;5% n (%)</td>
<td>199 (48.7)</td>
<td>186 (35.7)</td>
<td>187 (36.8)</td>
<td>87 (33.4)</td>
<td>101 (34.1)</td>
<td>168 (34.7)</td>
<td>73 (27.2)</td>
<td>110 (32.7)</td>
<td>132 (34.1)</td>
<td>1766</td>
</tr>
<tr>
<td>Functioning distractors per test n (%)</td>
<td>145 (32.9)</td>
<td>263 (50.5)</td>
<td>251 (50.0)</td>
<td>140 (53.8)</td>
<td>178 (60.1)</td>
<td>246 (54.0)</td>
<td>121 (45.1)</td>
<td>202 (60.1)</td>
<td>220 (58.0)</td>
<td>1766</td>
</tr>
<tr>
<td>Functioning distractors per item M (SD)</td>
<td>1.32 (1.14)</td>
<td>2.02 (1.20)</td>
<td>2.00 (1.21)</td>
<td>2.15 (1.19)</td>
<td>2.41 (1.01)</td>
<td>2.20 (1.22)</td>
<td>1.81 (1.14)</td>
<td>2.40 (1.21)</td>
<td>2.32 (1.20)</td>
<td>1766</td>
</tr>
<tr>
<td>Functioning distractors per item n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>31 (28.2)</td>
<td>14 (10.7)</td>
<td>16 (12.6)</td>
<td>7 (10.7)</td>
<td>3 (4.0)</td>
<td>10 (7.4)</td>
<td>10 (14.9)</td>
<td>6 (7.1)</td>
<td>7 (12.6)</td>
<td>104 (12.6)</td>
</tr>
<tr>
<td>One</td>
<td>37 (33.6)</td>
<td>35 (26.9)</td>
<td>31 (24.1)</td>
<td>11 (16.9)</td>
<td>8 (10.8)</td>
<td>25 (18.8)</td>
<td>18 (26.8)</td>
<td>15 (17.8)</td>
<td>16 (16.8)</td>
<td>96 (11.7)</td>
</tr>
<tr>
<td>Two</td>
<td>23 (20.9)</td>
<td>29 (22.3)</td>
<td>36 (28.3)</td>
<td>21 (32.3)</td>
<td>30 (40.5)</td>
<td>38 (33.3)</td>
<td>17 (25.3)</td>
<td>20 (23.8)</td>
<td>33 (34.7)</td>
<td>247 (27.5)</td>
</tr>
<tr>
<td>Three</td>
<td>14 (12.7)</td>
<td>38 (29.2)</td>
<td>28 (22.0)</td>
<td>17 (26.15)</td>
<td>22 (29.73)</td>
<td>19 (16.67)</td>
<td>19 (28.36)</td>
<td>25 (29.76)</td>
<td>18 (18.9)</td>
<td>200 (22.9)</td>
</tr>
<tr>
<td>Four</td>
<td>5 (4.5)</td>
<td>14 (10.7)</td>
<td>16 (12.6)</td>
<td>9 (13.85)</td>
<td>11 (14.86)</td>
<td>22 (19.30)</td>
<td>3 (4.48)</td>
<td>18 (21.43)</td>
<td>21 (22.1)</td>
<td>119 (12.4)</td>
</tr>
</tbody>
</table>
all four functional distractors from 2008 to 2010, with 7.6% items in the 2008 tests compared to 15.7% in 2009 and 18% in 2010. There were 25.8%, 18.5% and 21.1% of items with one functional distractor in 2008, 2009 and 2010, respectively. The number of items with two functional distractors increased in 2009 and 2010 (28.9% and 32.1%) compared to 2008 (26.2%), a similar trend was observed for items with three functional distractors except in 2010. Overall, the pattern was for increasing numbers of functional distractors per item from 2008 to 2010 after the introduction of peer review processes.

Discussion
This study highlights sustained, positive impacts on MCQ quality arising from the introduction of peer review processes into the development of MCQ items for examinations. The peer review process improved the quality of MCQ items based upon their utilisation in examinations.

Improvements in examination and item quality
The medical course uses graded assessments and so the reference ranges (DIF = 20–80%; DI = 0.16–1.00) that have been used to interpret the item analysis data for this study were considered appropriate. Results from this study show that the comparative proportions of easy, difficult, recall, non-discriminating and non-functional items were consistently higher pre-QA (2008) than post-QA (2009 and 2010). This confirms the report by Vahalia et al. (1995), that there is considerable evidence that MCQs are poorly written and too frequently assess recall rather than higher-order thinking. Poor quality items can have detrimental effects on learning and reliability and validity of assessment (Downing 2005). The opportunity to discuss and critically review questions enabled the faculty members to demonstrate their knowledge of student learning in their discipline. This was evident in the type of questions chosen to assess learning outcomes (e.g. application vs. recall) and also in the selection and rationale for the use of distractors in the MCQ.

A high quality MCQ is one in which each distractor is selected by at least some students who do not know the content tested by the question. Diagnosis of the degree of knowledge and the nature of an examinee’s misunderstanding can be made by assessing the functionality of distractors (Nitko 2004; Popham 2000). Given that assessment drives learning, a distractor that fails to attract any examinees is dysfunctional, does not assist in the measuring of educational outcomes, adds nothing to the item or the test (psychometrically) and has negative impact upon learners. In pre-QA implementation, 21% of the distractors used in the exams were implausible as compared to 10% afterwards. Tarrant, Ware, and Mohammed (2009) observed a similar pattern in their study. The low number of items with four functional distractors indicates that it is difficult to provide four plausible distractors for an in-house five-option set of MCQs. Research suggests that even professionally developed test items on standardised exams rarely have more than two functional distractors (Haladyna and Downing 1993). The use of MCQ with four functional distractors is seen as a ‘gold standard’ and was achieved with up to 22% of items in one examination in this study. However, others (Haladyna, Downing, and
Rodriguez 2002; Tarrant et al. 2009) suggest that achieving MCQ with three functional distractors is more realistic and practical.

**Examiner training and peer review**

Introduction of the peer review process occurred in a two-stage process operating in an open discussion format, facilitated by the authors. Whilst the faculty members recognised the importance and the need for the process (Malau-Aduli et al. 2011), introducing the process still met some resistance. The first stage focused on team building and standardising MCQ formats. The initial workshops focused primarily on the item structure, wordings and on the consistency of formatting: meetings were non-threatening, congenial, but longer than necessary and focused on minutiae rather than the content of question. The team themselves identified the need to move to more substantive discussions about the question content. Hence, the second phase introduced more structure to the meetings and focused on the content of questions. Question writers were explicitly requested to discuss (i) what the question tested and its structure (i.e. rationale for distractors); (ii) why the question was relevant or important; and (iii) if it tested core knowledge. This second phase provided more intellectual challenge, expression of personal disciplinary identity and expertise, and greater ownership and acceptance by the faculty. It also fostered a rigorous, but supportive environment for academic discussion about assessment and their discipline.

Using item analysis for school-based examinations revealed important and useful information that demonstrates that providing training for item writers improves the quality of the multiple-choice questions that they produce. The high number of non-discriminating items emphasises the need for QA processes such as judgmental analysis by subject-matter specialists and generalists to be set in place for improved quality of assessment items before they are administered (particularly for summative examinations). This study underscores the report of Josefowicz et al. (2002). They stated that to ensure the validity and reliability of test items; faculty members responsible for writing examinations should be trained because writing good examination questions is a skill that can be learned. This study confirms the fundamental benefit of using classical and summary test score analyses as basic tools to improve assessment. Providing item writers with timely and meaningful reports on their assessment items has fostered collegial engagement in best assessment practice by faculty.

Our results are in tandem with the findings of Wallach et al. (2006), who demonstrated that pre-establishing item-writing guidelines and peer review of exam questions by an interdisciplinary committee (before exam questions are administered to examinees) can greatly improve the quality of in-house examinations. Setting and notifying faculty members of committee review dates well in advance promoted better preparation of test items. The review process increased content validity of test items, reduced technical flaws and improved overall item quality because ambiguity in question interpretation was detected, multiple answers or lack of a one-best-answer were picked up, items tested higher cognitive learning abilities and suggestions were proffered to strengthen the test items before they were administered to the students.

The observed improvement in the quality of the items in this study has occurred in the context of an overall improvement of discussions and QA processes that included item-writing training workshops with established guidelines and peer
review of items. The QA processes were broad, but the only item quality focus was through the peer review. The peer review process improved the quality of questions’ content through the increased use of vignettes which require students to apply knowledge rather than recall facts. This was evidenced in the results from the classification of assessment items which showed a decrease in the proportion of questions rated as recall items post-QA compared to pre-QA (30% vs. 65%). The most crucial validity standard for a test item is that whether an examinee got an item correct or not is due to their level of knowledge or ability and not due to chances or test bias. The observed increased item statistics in this study post-QA indicated that the quality of the assessment items had improved and the tests were more accurate, valid and reliable.

Limitations
Generalisability of the findings from this study may be limited by several factors. This study has examined the effect of peer review on the quality of MCQ items over a period of three years of study and confirmed other research which have shown that pre-establishing item-writing guidelines and peer review of exam questions by an interdisciplinary committee (before exam questions are administered to examinees) can greatly improve the quality of in-house examinations (Josefowicz et al. 2002; Wallach et al. 2006). However, the observed outcomes in this study may not reflect possible outcomes from other medical programmes as some of the statistical improvements may have been due to measurement errors that have not been accounted for. This study also did not utilise external content experts to review the quality of the items, further studies could explore this and compare the results of both methods of validation of improvement.

Conclusion
Peer review was included as part of the process of MCQ item writing and examination development. The introduction of peer review led to sustained improvements in the quality of MCQ, specifically (i) questions written at the appropriate level, (ii) increased proportion of questions on examinations that discriminated between learners; and (iii) a decrease in the number of non-functional distractors and increase in functional distractors within each MCQ item.

Notes on contributors
Bunmi S. Malau-Aduli is lecturer in medical education (assessment), School of Medicine, University of Tasmania. Her research areas include the use of quality assurance processes to improve learning and teaching in higher education and the design and analysis of assessment instruments.

Craig Zimitat is deputy head of school and director of medical education, School of Medicine, University of Tasmania. His interests include teaching, learning and assessment in higher education and the use of technology to support academic work and student learning.

References


Downing, S.M. 2005. 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* 10, no. 2: 133–43.


