QUALITY ASSURING MULTIPLE-CHOICE QUESTION ASSESSMENT IN HIGHER EDUCATION

H. van der Merwe
Department of Educational Leadership and Management
University of South Africa
Pretoria, South Africa
e-mail: vdmerhm@unisa.ac.za

ABSTRACT

A growing scholarship links quality-assured multiple-choice testing to accountable outputs. This article looks at the use of multiple choice assessment and the quality assuring of item development through a structured process of systematic steps. A qualitative investigation was undertaken based on individual e-interviews with participants who are experts in the use of multiple-choice question assessment. The investigation was conducted at the College of Education of a higher education institution. The findings confirm the potential of multiple-choice assessment to test factual knowledge and higher-order learning. The findings also show that the main components in a systematic process of quality assuring the construction of multiple-choice items include training in the skills of item development, peer reviewing of constructed items, professional editing, and the interpretation of statistical analysis of student performance and student feedback for future constructions. The findings contribute to literature that argues for credible assessment practices to ascertain relevant outputs.

UNISA university of south africa

South African Journal of Higher Education
Volume 29 | Number 2 | 2015
pp. 279–297

ISSN 1011-3487

Keywords: assessment, quality assurance, deep learning, higher- and lower-order cognition, multiple-choice item construction, functional distractors

INTRODUCTION

In higher education, as is the case in all educational settings, tuition is focused on inputs to transform student learning through process-related arrangements into viable outputs. These outputs, based on specified tuition goals, are aimed at gained competencies (Nelson 2007, 24). To ascertain student learning, accountability systems based on quality-assured assessment practices are crucial. These systems represent a multi-level approach. At the institutional level, policies, guidelines and procedures direct the framework for assuring the integrity of the assessment system. At the school and module level, the balancing of formative and summative assessment strategies, the choice of instruments, staff training, and feedback processes contribute to the quality of assessment. At the individual level, personal choices with the selection of subject matter, engagement in training and review, and overall commitment to the assessment process affect its quality (Malau-Aduli and Zimitat 2012, 920). At all three levels, the quality assurance of assessment practices promotes the application of a systematic process in which a feedback loop is incorporated to prevent errors in ascertaining gained competencies.

Multiple-choice tests as a form of assessment supplementing other ways of assessment are widely used in higher education practices. Prompt feedback with low impact on marking time, accompanied by the advantages of computer networks for flexibility of time and location of testing, makes multiple-choice assessment a popular option for both the lecturer and the student (Nicol 2007, 54; Shavelson 2007, 30). If the test is long enough, a wide range of topics can be covered in one test with high reliability (Bush 2006, 398). In an environment of open access for improved egalitarianism, multiple-choice testing is a logical response to large student numbers and limited time for assessment due to modularisation combined with a semester system (O'Rourke 2009, 7; Singh and De Villiers 2012, 129). Multiple-choice testing is also a fair response to heterogeneous student compositions where criterion-based assessment incorporates student identity-development and student agency for alternative trajectory choices (Ghorpade and Lackritz 1998, 464; Handley, Den Outer and Price 2013, 891). With teaching and learning understood as a deliberative encounter in service of the public good (Waghid 2013, 1052), multiple-choice testing answers to self-criticality based on open engagement among lecturers and students.

But for multiple-choice assessment to be a deliberative encounter reflecting the nuanced levels of obtained student competency reliably, a sound process of quality assuring item development is important. Criticism against multiple-choice assessment as the mere testing of recall learning, and the thread to reliability and validity of badly written items (Malau-Aduli and Zimitat 2012, 920; Shavelson 2007, 14) amplifies

the importance of quality assurance for proper depth with the breadth of assessment. In this regard comprehensive research has been conducted on the aims, functions and procedures with higher education assessment and the possibilities of multiple-choice testing for different levels of student learning (Dickenson 2012; Fellenz 2004; Nicol 2007; Nicol and Macfarlane-Dick 2006; Shavelson 2007; Shavelson and Huang 2003; Singh and De Villiers 2012). Research has also been conducted on the quality assurance of multiple-choice tests and item construction (Bush 2006; Downing 2002; Malau-Aduli and Zimitat 2012; Wallach et al. 2006). What has been less researched are the context-specific measures employed to quality assure item development for the comprehensive testing of obtained competency. The rigour with item construction could counter limitations and enhance the relevance of multiple-choice testing as an assessment tool within the specific context.

This article focuses on the use of multiple-choice assessment and the way of quality assuring item construction at module level in a higher education context. In light of the prominence of multiple-choice assessment, I argue that the rigour of a systematic process of quality assuring multiple-choice item development at module level can contribute to the discourse on improved assessment practice to ascertain relevant outputs.

The point of departure in this article is the theories of Bloom and Gardner on cognitive learning and multiple intelligences as a theoretical framework underlying the qualitative investigation based on individual e-interviewing of 28 purposefully selected participants. Findings from an analysis of the data are followed by a discussion of the purpose of using multiple-choice assessment, the aspects pertaining to setting acceptable items and a systematic process of quality assuring item construction.

THEORETICAL AND CONCEPTUAL FRAMEWORK

The theoretical framework underlying the focus on quality-assuring multiple-choice item construction is supported by a clarification of the features of a good quality multiple-choice question, with advantages and disadvantages of multiple-choice assessment. A reflection on the learning goals and outcomes in higher education tuition is included to ascertain what is implied with testing comprehensively to ensure that outputs reflect what ought to have been attained.

The theories of Bloom and Gardner on cognitive learning and multiple intelligences

Bloom's taxonomy of learning, representing a hierarchy that classifies cognitive processes from lower-order remembering to higher-order critical and creative thinking, is relevant to debates on the appropriateness of multiple-choice questions to test the different levels of learning. The normal thinking process represents

remembering before one is able to understand knowledge in order to proceed to application and eventual creative engagement (Anderson and Krathwohl 2001, 58). The associated learning activities representing a hierarchical order of lower- to higher-order cognitive engagement include the following (Anderson and Krathwohl 2001, 67):

- Remembering: recognising, listing, describing, identifying, naming, locating
- Understanding: interpreting, summarising, classifying, comparing, explaining
- Applying: implementing, carrying out, using, executing
- Analysing: comparing, organising, attributing, outlining, structuring, integrating
- Evaluating: checking, critiquing, experimenting, judging, testing, detecting
- Creating: designing, constructing, planning, producing, inventing, devising

When assessment instruments, including multiple-choice questions, are developed to assess learning outcomes, the compiler should have knowledge of these categories of cognition, from simple to complex, and of the learning activities associated with the specific taxonomic element.

Gardner's identification of multiple intelligences that emphasise different strengths in humans and different ways of learning expands the traditional view on intelligence. In addition to logical-mathematical and linguistic intelligences, the existence of spatial, musical, bodily-kinaesthetic, interpersonal, and intrapersonal intelligences broadens the perception of human potential and learning (Blythe and Gardner 1990). The broader concept of knowledge, skills and behaviour gained with learning influences the assessment of that learning. Multiple intelligences also appeal for sensitivity regarding differences in valued intelligences across societies. Particular intelligences might be highly evolved in many people of one culture, but less developed in the individuals of another society (Fellenz 2004, 705; Noble 2004, 197). The wider concept of intellectual strength appeals to teaching and learning and the assessment of that learning to be a deliberative encounter between student and lecturer based on accommodation and social transaction that is constituted by openness, trust and self-criticality (Waghid 2013, 1051).

Considering the theories of multiple intelligences alongside those of levels of learning, students generally may engage higher-order thinking and problem solving in an area of intellectual strength and only lower-order thinking in an area of relative weakness. The challenge this holds for tuition with accompanying assessment, which includes multiple-choice assessment, is the testing of student learning through different intellectual domains at the same or different levels of cognitive complexity (Fellenz 2004, 705; Noble 2004, 194). The testing of multiple intelligences and multiple levels of cognition encapsulates the potential of rhizomatic tuition to escape learning of a linear and unidirectional format to transcend to what Waghid (2013, 1053) identifies as 'new assemblages of meaning'. Competency gain through the

testing of rhizomatic learning is naturally enhanced when students do challenging higher-order tasks in the intellectual domain that they find comfortable.

GOALS AND OUTCOMES IN HIGHER EDUCATION LEARNING

Typical of the business approach to which higher education institutions are exposed is the culture of evidence that is demanded with all aspects of institutional functioning (Hoecht 2006; Nelson 2007; Van Wyk 2005). This evidence includes accounting for student learning by reporting on the extent to which students are learning, or have learned, what was intended. Assessing what students are learning through formative assessment, accompanied by feedback that is based on openness, trust and self-criticality, improves students' quality of learning experiences, and their autonomy in learning (Nicol and Macfarlane-Dick 2006, 201; Waghid 2013, 1051). Learning autonomy is especially relevant in open distance learning environments with less personal contact between students and teaching staff (O'Rourke 2009, 8). Determining what students have learned through summative assessment engenders eventual qualification based on high-stakes judgements of students' competency gain (Nicol 2007; Nicol and Macfarlane-Dick 2006; Noble 2004). However, evidence of learning does not imply competency gain if it is not related to deep learning (Shavelson 2007, 28; Van Wyk 2005, 8). Apart from student learning as a deliberative encounter based on a student-centred and problem-solving approach (Waghid 2013, 1051), tuition goals in higher education pertain to acquiring knowledge, skills and behaviour of a cognitive, personal, social and civic nature within a specific field of study (Shavelson and Huang 2003, 12). To ensure deep learning, and avoid surface learning that is linked to superficial assessment (McLachlan 2006, 716), these tuition goals should be tested comprehensively.

Learning outcomes in higher education depart from discipline-based knowledge with application possibilities to broader and different contexts. Based on multiple intelligences theory, what is learned depends on the competencies students gained from prior exposure and their own natural abilities and interests. When learning, students draw on the identity trajectories available in their communities, or the ones they are joining, but with the built-in agency of developing alternative trajectories to imagine things anew (Handley et al. 2013, 891; Waghid 2013, 1052). Considering student learning comprehensively, cognitive outcomes entail the skills of remembering in order to gain knowledge of particular aspects in a specific discipline and to think critically so as to solve problems and communicate clearly (Shavelson 2007, 26). Personal and social outcomes involve the competencies of perspective taking, experiencing empathy, having consideration, and developing self-comprehension. Perspective taking to consider the world from different viewpoints, together with empathy to allow for congruent emotions, capacitates students in comprehending the self and others (Forgas, Kruglanski and Williams 2011, 104). Civic outcomes

entail the ability to balance personal and social goals and to take initiative for social responsibility by maintaining normative conditions through coordinating efforts with others and interacting with people from various backgrounds (Louw 2010, 48).

What should be clear is that the cognitive outcomes achieved with higher education learning are riddled with 'soft skills', such as, for example, reasoning that is accompanied by personal relations, moral challenges and civic engagement, which, in a context of communicative openness, engender the collective worth of human beings. As what is taught to students is also assessed, 'soft skills' should naturally be included in assessment practices to ensure that outcomes reflect deep learning (Nelson 2007, 26; Shavelson 2007, 33). Further, for outputs to adhere to the criteria of deep learning, assessment practices should be based on students demonstrating their gained competencies through different intellectual domains at the same or different levels of cognitive complexity, with the accommodation of student agency for new ways of thinking (Anderson and Krathwohl 2001; Fellenz 2004; Handley et al. 2013; Noble 2004; Waghid 2013).

USING MULTIPLE-CHOICE QUESTION ASSESSMENT

Multiple-choice testing has many advantages. The efficacy in terms of time to assess a large number of students and the opportunity of assessing objectively without having to deal with the subjective interpretation of different markers are important benefits (Costagliola and Fuccella 2009, 82; Dickenson 2012, 2). Assessment is carried out within a short period of time as the approach is the selection instead of the creation of answers, which enables a comprehensive evaluation of the extent of knowledge (Pitenger and Lounsbery 2011, 8). Efficiency in terms of time and content coverage is enhanced by using online testing possibilities. Factors irrelevant to the assessed content, such as handwriting, do not influence answering and students are graded purely on their knowledge of the learning content (Fellenz 2004, 704). Students are not graded on their ability to express themselves in English, which enhances equal opportunity for progress crucial to a heterogeneous student corps consisting of second language speakers from different 'ethnic groups' (Bush 2006, 398; Ghorpade and Lackritz 1998, 464). The onus of constructing items in clear and simple English for optimal understanding, which can be ensured with proper quality assurance, is obviously very important. If properly constructed, multiple-choice questions can test higher-order learning and can accurately discriminate between high- and low-achieving students (Tarrant, Ware and Mohammed 2009, 1).

Limitations to multiple-choice testing relate to item development that is time-consuming and demanding. To design items with functioning distractors that encapsulate different levels of cognition is challenging (Chiheb, Faizi and Afia 2011, 70; Tarrant et al. 2009, 4). Items with questionable quality due to dysfunctional distractors affect the validity and reliability of the assessment (Dickenson 2012, 1;

Nicol 2007, 54). The reason is that dysfunctional distractors, understood as options that are selected infrequently, enable students to select the right answer regardless of their knowledge of the topic (Downing 2002, 236; Malau-Aduli and Zimitat 2012, 921). Dysfunctional distractors therefore do not contribute to the testing of learning outcomes, but have a negative impact on the question, the test and student morale (Tarrant et al. 2009, 2).

As lecturer perspectives are dominant, a student's answer may elicit an 'incorrect' response when the student fails to interpret information as the lecturer intended. This negates the existence of pluralities and relativities, which jeopardises a learning-centred approach to tuition (Fellenz 2004, 704). The fairness of multiple-choice testing for heterogeneous student populations demands thorough knowledge of the interests and aptitudes of the differentiated student corps to avoid item bias (Downing 2002, 238). As multiple-choice testing is based on selection only, the result may be that even if students have some knowledge of the content, they receive no credit for knowing that information if they select the wrong answer (Pitenger and Lounsbery 2011, 5). Similarly, random guessing leading to correct answer selection results in a construct-irrelevant variance that reduces the validity and reliability of test score interpretation (Downing 2002, 236). This can, however, be negated by a system of negative marking that encourages students to associate confidence with each of their selected answers (Bush 2006, 402; Ventouras et al. 2011, 619).

Some researchers argue that multiple-choice tests promote memorisation and factual recall without the possibility of assessing higher-order cognitive processes (Nelson 2007; Chiheb et al. 2011; Ventouras et al. 2011). Others maintain that it depends on the approach followed with item construction and that multiple-choice questions can be functionally used to test learning at higher-order levels of cognition including analysis, evaluation and creativity (Bush 2006; Fellenz 2004; Nicol 2007). Further, when quality assurance is effectively applied with construction, the possibility exists for an integration of the testing of higher-order levels of learning with the possibility of accommodating multiple intelligences resulting in increased egalitarianism (Noble 2004, 202; O'Rourke 2009, 6).

QUALITY ASSURANCE OF MULTIPLE-CHOICE ITEM CONSTRUCTION

The quality of tuition is directly linked to the quality of the assessment of that tuition. Quality assurance in the construction of multiple-choice items should represent a systematic process of specific actions to ensure that assessment requirements are met and that 'defects' are removed from the final 'product'. These actions can be grouped into pre-test and post-test quality assurance initiatives undertaken in a peer review setting (Bush 2006; Malau-Aduli and Zimitat 2012). Pre-test quality assurance actions include matters such as adopting best practices regarding item

format and peer reviewing of constructed questions. Post-test quality assurance measures represent statistical analysis of the conducted test and considering student feedback (Bush 2006, 399).

A starting point for item construction as part of the peer reviewed pre-test quality assurance initiative is the consideration of Bloom's taxonomy of different cognitive levels of learning and testing, Gardner's classification of multiple intelligences, and knowledge of the proper format of a multiple-choice item (Fellenz 2004, 705; Noble 2004, 194). Considering item writing guidelines to support the writing of multiple-choice items improves its quality (Wallach et al. 2006, 66). Reference to item-writing guidelines with item construction should be accompanied by a reflection on the choice of correct answer, the reason for the incorrectness of each distractor, and the level and type of tested learning (Fellenz 2004, 706; Noble 2004, 195).

Feedback on construction in a peer reviewed setting focuses on the quality of item design, the accuracy and justification of the correct and incorrect answer options, the identification of the tested cognitive level and intellectual domain, and the overall suitability of the developed items for assessing course-relevant learning (Fellenz 2004, 707). Criteria for course-relevant learning relate to what the item is testing, why the item is important, and whether core knowledge is tested (Malau-Aduli and Zimitat 2012, 923). Editing as an important quality assurance measure to improve English language and terminology enhances item performance. An important aspect of editing is to eliminate uncommon words, especially when second-language speakers are writing the test (Bush 2006, 401). The value of peer reviewed feedback on item construction is that the purpose of assessment in terms of goals, criteria and standards with testing is clarified (Nicol 2007, 55). Reading item constructions out loud within a peer review setting enhances the possibility of detecting ambiguity in question interpretation (Wallach et al. 2006, 66). Peer reviewed feedback promotes self-assessment, which enhances self-efficacy and ownership-taking, which in turn promote an overall improved conduct with multiple-choice question assessment (Nicol 2007, 59).

With regard to post-test quality assurance, aspects to consider include statistics on the number of examinees, the highest and lowest scores, median and mean averages, standard deviation and distractor functionality (Bush 2006; Tarrant et al. 2009). Two matters to consider with each multiple-choice item are its difficulty in terms of the proportion of examinees who have answered it correctly and its discrimination value in terms of the correlation of question performance to test performance (Bush 2006, 403). Further, statistics on each distractor determines the quality of the question. With a high quality multiple-choice question, each distractor is selected by at least some students who do not know the content tested by the question (Malau-Aduli and Zimitat 2012, 927).

A very important aspect of post-test quality assurance of multiple-choice testing is the gaining and concerted consideration of student feedback on the test they have

written. Apart from achieving process improvement, the consideration of student feedback as a deliberative encounter based on openness, mutual trust and critical engagement promotes rhizomatic tuition and assessment engendering improved egalitarianism (Bush 2006; Ghorpade and Lackritz 1998; Waghid 2013).

RESEARCH DESIGN FOR THE EMPIRICAL INVESTIGATION

The empirical investigation into the practice of multiple-choice assessment and the measures of quality assuring item construction was conducted in response to the appeal from the authorities of a higher education institution to conduct such an investigation at all the different colleges of the institution. In line with the work of Nelson (2007, 24), the appeal was motivated by the universal demand for institutional accountability with assessment, namely ensuring that their own practice adheres to an alignment of predetermined outcomes with outputs. The investigation on which this article is based was conducted at the College of Education at the specific institution. The aim of the investigation was to understand the purpose of the particular use of multiple-choice question assessment, the features considered characteristic of a good item, and the application of quality assurance arrangements with construction. A related aim pertained to the development of a process for quality assuring multiple-choice item construction in pursuit of good practice.

The investigation took place within an interpretive paradigm using a qualitative research approach employing e-interviewing. In line with the experiences of Bampton and Cowton (2002), e-interviewing was asynchronously conducted, giving the interviewees time to reflect before supplying considered replies. In many instances the initial e-interviews resulted in further prompting through interactive communication via e-mail for the sake of a deeper understanding of the process of quality assuring the construction of proper multiple-choice items.

Teaching staff with 15 years or more of teaching experience in a higher education environment and who acted as primary lecturers of modules in which multiple-choice tests were used were approached. A total of 28 participants took part in the investigation. In line with the findings of Henning, Van Rensburg and Smit (2004) and Rossman and Rallis (2003), this total was considered sufficient as information-rich participants were selected. Criteria for information-rich selection were based on the indicators of rank and years of experience. Rank served as an indicator of authority based on work-related knowledge and skills whereas years of teaching experience confirmed acquired competencies due to prolonged engagement in the specific field of practice (Cohen, Manion and Morrison 2011, 88). The inclusion of less experienced staff could have revealed extended nuances on the engagement with multiple-choice question practice and the development of a process of quality-assuring the construction of items. However, the focus on experienced staff served

the purpose of comparing universally applied actions with multiple-choice item construction with applications by expert staff within the specific context, with as final aim the refining of own practice.

Participants were individually approached by means of e-interviewing to share their knowledge, skills and conduct on multiple-choice question assessment. They were asked to respond to three themes on multiple-choice question practice, namely the purpose of use, indicators for best practice, and measures of arranging for quality assurance with item construction. Participants were prompted to share any additional knowledge on the improvement of multiple-choice question practice.

Qualitative content analysis was done based on Tesch's model to ensure that all the perspectives and issues arising from the data were included in the report (De Vos 2005, 337). In brief, this meant that each interview was read for an immersion into the data and as an initial segmentation of the data into units of meaning. This step was followed up with open coding by reading and re-reading each interview to ensure an overview of as much contextual data as possible, so as to achieve an inductive selection of codes determined at sentence level (Henning et al. 2004, 104). After axial coding was done, selective coding ensured that themes from the labelled categories were constructed and extracted to represent the interpreted and rationalised data as research findings (Henning et al. 2004, 105). Guba's model on trustworthiness, as explained by De Vos (2005, 346–347), was applied to ensure the authenticity of findings in terms of truth value relating to data-checking by participants, consistency in the form of an audit trail, applicability to other similar situations as confirmed by the researchers of those situations, and neutrality as the unfolding revelation of a studied reality. Apart from member checking, the research findings from the empirical investigation were triangulated with the findings from a review of the literature. The anonymity of participants and the confidentiality of their disclosures were guaranteed at all times during the research project.

FINDINGS

Multiple-choice question practice is discussed through three themes. These themes, relating to the themes posed with e-interviewing concurring with what was identified in the literature, pertain to the purpose of multiple-choice question use, features of good items, and quality assurance measures with the construction of items. These themes are discussed below and substantiated by verbatim excerpts from the interviews.

The purpose of multiple-choice question use

With the majority of modules offered at the research site, the first assignment of the year, which represented formative assessment, consisted of multiple-choice questions. In line with the requirements of the National Department of Education

that higher education institutions should present data and statistics on active students, this first assignment served a subsidy-securing purpose. This first assignment also served as examination admission. From the participants' point of view, the purpose with this first, formative assignment was to guide students to an overview of the module content with inclusion of items that promoted higher-order learning.

Lecturer participants agreed that the multiple-choice question assignment, as the first formative assessment of the year, served the following purposes:

- 'a "testing" of the students' understanding of the content'
- 'to force students to go through their work at an early stage'
- 'to increase content coverage while testing higher-order learning'
- 'to prevent students from ignoring topics, which they do not expect to be tested on'

Participants emphasised that since essay-type assessment allows a limited number of questions, content coverage is limited, which gives rise to undesirable study strategies. Good quality multiple-choice assessment counteracts the possibility of students completing their assignments and passing their examinations after studying only part of the module content. As emphasised by a participant, 'potential gaps in the knowledge base of students may fail to become manifested' with the use of essay-type questioning only.

Some participants perceived multiple-choice questions to be mainly used to assess lower-order learning relating to 'factual knowledge' and the testing of 'recognition and understanding'. Others shared the opinion that the questions are also applied to test higher-order cognitive competencies such as 'comparing', 'critiquing' and 'producing'. These participants emphasised that although the construction of items to test higher-order learning are time-consuming, once constructed and 'proved to be valid', these items represent an important part of the item bank for testing deep learning. Since the research site reflected a heterogeneous student corps, participants acknowledged the potential of multiple-choice items to accommodate the testing of multiple intelligences. However, the prevailing attitude displayed a lack of knowledge on application possibilities: 'I know this [accommodation of multiple intelligences with testing] is possible, but don't ask me how', which emphasised the need for professional training in multiple intelligences accommodation with the construction of multiple-choice questions.

Participants acknowledged the benefit of time cost with multiple-choice assessment, which represented a section of their examination papers 'to reduce the marking load which would otherwise be unbearable'. Participants involved in semester courses emphasised that they did not have any other option than to use multiple-choice assessment only as they 'can't mark essay-type scripts of such magnitude on time'. Multiple-choice question assessment was also functionally

used to prevent the substantial marking variations prevailing with external marker assistance. Since the institution had almost 400 000 registered students, a considerable number of lecturers and external markers were involved in the marking of essay-type assignments and examinations of modules comprising large student numbers. In some modules 'six external markers assist with the marking of the assignments and examination papers' due to the magnitude of student numbers. Lecturers' and external markers' 'widely differing backgrounds and interpretations' often result in substantial marking variations that impact negatively on the fairness of assessment. Such a situation was prevented by multiple-choice assessment 'with marking done objectively by the computer'.

It was clear that multiple-choice assessment was widely used at the research site due to the time constraints associated with a semester system with large student numbers, and to counteract unfair assessment practices with subjective marking variations. Participants understood multiple-choice assessment as being readily suitable for assessing lower-order cognitive skills; however, with ample possibilities to test higher-order learning. Although multiple-choice assessment extended the testing of factual module content to incorporate higher-order learning, it became evident that know-how on the accommodation of multiple intelligences with assessment initiatives was, to a large extent, still lacking.

Features of good items

As experienced lecturers were approached for their opinions on multiple-choice practice, their suggestions related to what is commonly known as prerequisites for setting good quality multiple-choice items. Aspects such as to 'use plausible distractors' were emphasised. It was pointed out that plausibility in terms of the accommodation of multiple intelligences responses can be arranged by using the answers that students gave in previous essay-type examinations to provide realistic distractors. Apart from same-length distractors 'options should be homogenous in content' and 'the correct answer must not be longer or shorter than the distractors'. Participants emphasised that multiple-choice questions should be set as questions rather than incomplete statements, negative questions should be avoided, and simple, precise and unambiguous wording should be used. Participants felt strongly about the negative influence of poorly formulated questions on the reliability of multiple-choice question assessment as 'poorly formulated questions result in well-informed students answering them incorrectly'. Participants also pointed out that the question (stem) should contain only a single problem because 'if it contains more than one problem and many students fail to find the correct answer to the question, it will not be possible to identify which problem caused the error'. Participants also pointed out that distractors such as 'all of the above', or 'none of the above' should not be used because students either 'merely recognise two correct options to get the answer

correct', or, it is not clear whether students actually knew the correct answer as 'only their ability to detect incorrect answers were tested'.

Following the normal rules of grammar and punctuation was indicated as being basic to proper multiple-choice question formulation. Aspects that were identified included that distractors must begin with capital letters if the stem is in question form; distractors serving to complete the stem as a statement begin with lower-case letters. Full stops are not used with numerical distractors 'to avoid confusion with decimal points'. It was also emphasised that compilers should be alert to verbal clues that either eliminate a distractor, or that lead to the correct answer. Participants mentioned examples of verbal clues such as similarity of wording in the stem and the correct answer, the correct answer including more textbook language than the distractors, and the use of absolute terms such as 'all', 'only' or 'never' 'which usually depicts a distractor'.

It was evident that participants were well versed in the development of plausible multiple-choice questions. They confirmed that they had undergone training in the construction of multiple-choice items at some stage during their career. They acknowledged this training as being critical to multiple-choice question practice in order to improve the quality and fairness of the assessment so as 'to reduce the number of student queries and arguments'.

A process of quality assuring the construction of items

At the research site, lecturing teams consisting of three to four lecturers per module share the lecturing responsibilities for the specific module, which include the setting of assignments (formative assessment) and examination papers (summative assessment). The conduct followed by the participants with their lecturing teams in constructing multiple-choice items was taken into consideration to develop a structured process of systematic steps according to which multiple-choice item construction can be applied to ensure the least possibility of 'defects' in the final 'product'. The process is diagrammatically depicted in Figure 1.

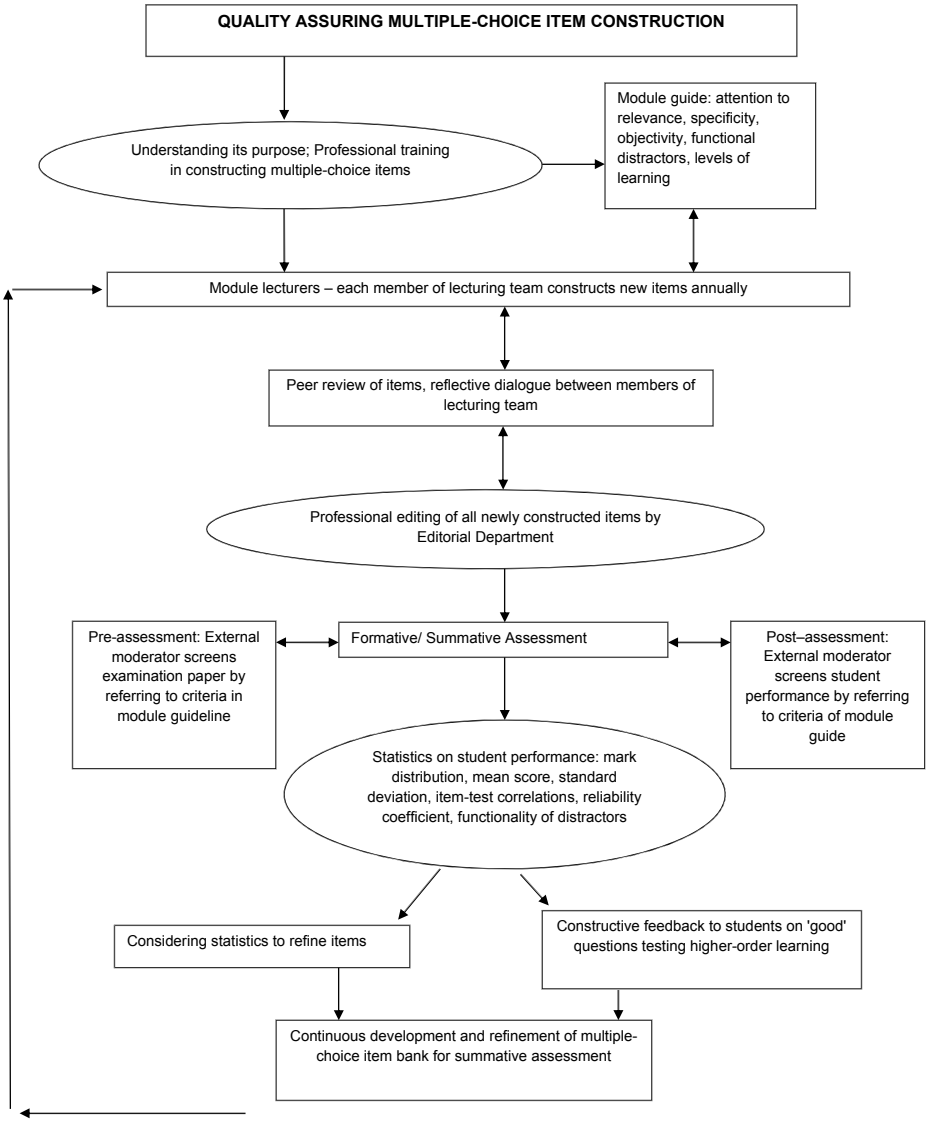

Figure 1: The process of quality assuring multiple-choice item construction

According to the process depicted in Figure 1, new multiple-choice items are constructed annually to serve the purpose of formative assessment while contributing to the existing item bank and eventual summative assessment. Module team members understand the purpose of the multiple-choice assessment as content coverage while also encouraging higher-order learning. Each member of the lecturing team drafts a predetermined number of new multiple-choice items per year. These items are

constructed according to criteria included in a module guide on proper multiple-choice question construction. The module guide, which was initially developed by members of the module lecturing team, is yearly refined with inputs from the existing lecturing team. These inputs are based on extended know-how gained from consulted literature and personal experience on item construction. Every member of the lecturing team has completed an initial professional training course in multiple-choice item construction.

Every set of newly compiled questions is reviewed individually by all the lecturers of the lecturing team. This review is followed by a team member meeting where compilers engage in reflective dialogue on each item construction. The reflection on each item is preceded by reading the item out loud. The criteria for reflecting on the validity and reliability of each constructed item pertain to the following:

- Relevance – the question is focused on core knowledge.
- Specificity – question construction directs answers to specific knowledge content.
- Objectivity – the degree of difficulty and discriminating power reflects the nature of the question as neither too easy nor too difficult to correlate with the competencies applicable to the module outcomes and to discriminate between learners' competency.
- Distractor functionality – each distractor is based on a common misconception about the correct answer.
- Level of testing – items represent the proportionate testing of lower- and higher-order learning.

The newly constructed, reviewed items are included as formative assessment in the first assignment submitted by the students.

The multiple-choice question assignment is marked via computer. Once marked, the lecturing team receives a statistical analysis of the results, which include information on derivatives, a distribution of test scores, the mean score and standard deviation, item-test correlations, reliability coefficients, and distractor functionality. By considering these statistics the lecturing team identify 'good and bad items'. Specific item deficiencies are determined and the possibility of amending deficiencies considered. 'Un-repairable, bad questions' are removed and 'all the good questions become part of the item bank'. The item bank is continuously extended and refined. For summative evaluation 'only item bank items are used'. The questions from the item bank are again scrutinised based on pre-test peer reviewing. This repeating of the pre-test evaluation of existing items implies that 'quality assurance of multiple-choice items remains an iterative process'. The external examiner of the module's summative assessment also evaluates the questions against the criteria for multiple-

choice question drafting included in the module's guide on multiple-choice question construction.

An aspect emphasised by all the participants as crucial to any process of quality assurance is the professional editing of all newly constructed multiple-choice items by the institution's editorial department. Participants pointed out that even when mother tongue speakers develop multiple-choice questions, editing is important as 'an objective reading of the question by an outsider contributes to optimal clarity', which is crucial to construct validity and the fairness of testing. Providing feedback to students with a special focus on 'good questions', which are considered 'challenging due to testing higher-order learning' is considered vital for proper learning and an important aspect of formative multiple-choice question assessment.

A gap determined in the systemised process of quality assuring multiple-choice item construction at the research site was the lack of a concerted effort to collect and interpret student feedback on testing. Examinees have worthwhile comments about the multiple-choice test they have written. Gaining students' feedback, based on a stance of openness, and concertedly considering this feedback with item development will provide opportunities for the accommodation of multiple intelligences that will result in improved fairness and increased egalitarianism with assessment. Concertedly incorporating student feedback into item development will also contribute to tuition and assessment becoming a deliberative encounter between student and lecturer based on mutual trust and critical engagement with the built-in potential of tuition and assessment becoming rhizomatic in nature.

CONCLUSION

Criticism against multiple-choice testing holds that this form of assessment is limited to the testing of factual knowledge at lower-order levels of cognition and that poor construction of items due to inadequate attention to the writing process affects the validity and reliability of the assessment. It is important to find ways of countering this criticism because multiple-choice testing is a relevant assessment practice due to its efficiency in terms of time spent on comprehensive content coverage and the fairness of student competency assessment not being hampered by difficulty with expression through English. In a higher education environment characterised by massive student numbers, multiple-choice assessment provides the additional advantage of contributing to the fair marking of vast responses. Although a strong predictor of students' factual knowledge, multiple-choice items can be constructed in such a way to incorporate testing at all levels of cognitive complexity with the possibility of accommodating multiple-intelligence responses. Reliability is, however, contingent on the rigour with which construction is quality assured.

Concurrent with the work of Bush (2006), Fellenz (2004), Malau-Aduli and Zimitat (2012) and Nicol (2007), the quality assurance of multiple-choice practice can be arranged through a systematic process of item construction with functionally interwoven peer reviewed feedback. Applying a systematic process helps to remove defects from the final multiple-choice assessment instrument. The main features of a structured process of item construction are:

- understanding the purpose of multiple-choice question assessment,
- professional training in the skill of item development,
- peer reviewing of constructed items,
- reflective dialogue on constructions,
- professional editing,
- consideration of objective input from external assessors, and
- interpretation of statistical analyses of item and student performance for refined reconstruction.

Quality assuring multiple-choice assessment through this process of systematic steps enhances accurate estimates of students' competence to confirm higher education learning outcomes accountably. It is suggested that further research be conducted on ways to combine the testing of higher-order learning with the testing of multiple intelligences. In higher education environments characterised by heterogeneous student populations, insights into the integration of higher-order learning and multiple intelligences can contribute to the discourse on accountable multiple-choice assessment for improved egalitarianism. For the sake of a complete and fair process of quality assuring item construction, it is also suggested that the gaining and interpretation of student feedback as a post-test endeavour be pursued concertedly. Inclusion of examinee feedback in the systemised process of item construction will contribute to improved validity and reliability with multiple-choice assessment. Concertedly considering student feedback will also contribute to tuition and learning becoming a deliberative encounter between student and lecturer. When based on openness and critical engagement, deliberative encounter could promote tuition and multiple-choice assessment becoming rhizomatic, thus representing deep learning and an improved practice with the assessment of gained competencies.

REFERENCES

Anderson, L. W. and D. Krathwohl. 2001. *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Addison, Wesley Longman.

Bampton, R. and C. J. Cowton. 2002. The e-interview. *Forum: Qualitative Social Research* 3 (2): 1–11.

Blythe, T. and H. Gardner. 1990. A school for all intelligences. *Educational Leadership* 47(7): 33–37.

Bush, M. E. 2006. Quality assurance of multiple-choice tests. *Quality Assurance in Education* 14(4): 398–404.

Chiheb, R., R. Faizi and A. E. Afia. 2011. Using objective online testing tools to assess students' learning: Potentials and limitations. *Journal of Theoretical and Applied Information Technology* 24(1): 69–72.

Cohen, L., L. Manion and K. Morrison. 2011. *Research Methods in Education*. (7th ed.). London: Routledge Falmer.

Costagliola, G. and Fuccella, V. 2009. Online testing, current issues and future trends. *Journal of e-Learning and Knowledge Society* 5(3): 79–90.

De Vos, A. S. 2005. Qualitative data analysis and interpretation. In *Research at grass roots. For the social sciences and human service professions,* ed. A. S. de Vos, 333–349. 3rd ed. Pretoria: Van Schaik.

Dickenson, M. 2012. The thing about multiple-choice tests. *Learning Solutions* June: 1–3.

Downing, S. M. 2002. Threats to the validity of locally developed multiple-choice tests in Medical Education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education* 7(3): 235–241.

Fellenz, M. R. 2004. Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment & Evaluation in Higher Education* 29(6): 703–719.

Forgas, J. P., A. W. Kruglanski and K. D. Williams. 2011. *The psychology of social conflict and aggression*. New York: Psychology Press.

Ghorpade, J. and J. R. Lackritz. 1998. Equal opportunities in the classroom: Test construction in a diversity-sensitive environment. *Journal of Management Education* 22(4): 452–471.

Handley, K., B. den Outer and M. Price. 2013. Learning to mark: Exemplars, dialogue and participation in assessment communities. *Higher Education Research & Development* 32(6): 888–900.

Henning, E., W. van Rensburg and B. Smit. 2004. *Finding your way in qualitative research*. Pretoria: Van Schaik.

Hoecht, A. 2006. Quality assurance in UK higher education: Issues of trust, control, professional autonomy and accountability. *Higher Education* 51(4): 541–563.

Lentell, H. 2007. Curriculum development – what is the role of ODL? *Proceedings of the African Conference on Higher Education*, September, Unisa, Pretoria.

Louw, W. 2010. Africanisation: A rich environment for active learning on a global platform. *Progressio* 32(1): 42–54.

Malau-Aduli, B. S. and C. Zimitat. 2012. Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education* 37(8): 919–931.

McLachlan, J. C. 2006. The relationship between assessment and learning. *Medical Education* 40(8): 716–717.

Nelson, C. 2007. Accountability. The commodification of the examined life. *Change* November/ December: 22–27.

Nicol, D. 2007. E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education* 31(1): 53–64.

Nicol, D. J. and D. Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education* 31(2): 199–218.

Noble, T. 2004. Integrating the revised Bloom's taxonomy with multiple intelligences: A planning tool for curriculum differentiation. *Teachers College Record* 106(1): 193–211.

O'Rourke, J. 2009. Meeting diverse learning needs. *Progressio* 31(1&2): 5–16.

Pitenger, A. L. and J. L. Lounsbery. 2011. Instructional design and assessment: Student-generated questions to assess learning in an online orientation to pharmacy courses. *American Journal of Pharmaceutical Education* 75(5): 1–8.

Rossman, G. B. and S. F. Rallis. 2003. *Learning in the field: An introduction to qualitative research.* 2nd ed. Thousand Oaks, CA: Sage.

Shavelson, R. J. 2007. Assessing student learning responsibly: From history to an audacious proposal. *Change* January/February: 26–33.

Shavelson, R. J. and L. Huang. 2003. Responding responsibly to the frenzy to assess learning in higher education. *Change* January/February: 11–19.

Singh, U. G. and M. R. de Villiers. 2012. The use of different types of multiple-choice questions in electronic assessment. *Progressio* 34(3): 125–143.

Tarrant, M., J. Ware and A. M. Mohammed. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education* 9(40): 1–8.

Van Wyk, B. 2005. Performativity in higher education transformation in South Africa. *South African Journal of Higher Education* 19(1): 5–19.

Ventouras, E., D. Triantis, P. Tsiakas and C. Stergiopoulos. 2011. Comparison of oral examination and electronic examination using paired multiple-choice questions. *Computers and Education* 56(3): 616–624.

Waghid, Y. 2013. Teaching and learning as a deliberative encounter: On the possibility of new imaginings. *Higher Education Research & Development* 32(6): 1051–1053.

Wallach, P. M., L. M. Crespo, K. Z. Holtman, R. M. Galbraith and D. B Swanson. 2006. Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education* 11(1): 61–68.