Nurse Education Today (2006) **xxx**, xxx–xxx

ELSEVIER

Nurse Education Today

intl.elsevierhealth.com/journals/nedt

The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments

Marie Tarrant [a,*], Aimee Knierim [b], Sasha K. Hayes [a], James Ware [b]

[a] *Department of Nursing Studies, Faculty of Medicine, University of Hong Kong, 4/F, William M.W. Mong Block, 21 Sassoon Road, Hong Kong*
[b] *Teaching and Learning Centre, Faculty of Medicine, Chinese University of Hong Kong, Hong Kong*

Accepted 20 July 2006

KEYWORDS
Multiple-choice questions;
Item-writing flaws;
Assessment;
Examination

Summary Multiple-choice questions are a common assessment method in nursing examinations. Few nurse educators, however, have formal preparation in constructing multiple-choice questions. Consequently, questions used in baccalaureate nursing assessments often contain item-writing flaws, or violations to accepted item-writing guidelines. In one nursing department, 2770 MCQs were collected from tests and examinations administered over a five-year period from 2001 to 2005. Questions were evaluated for 19 frequently occurring item-writing flaws, for cognitive level, for question source, and for the distribution of correct answers. Results show that almost half (46.2%) of the questions contained violations of item-writing guidelines and over 90% were written at low cognitive levels. Only a small proportion of questions were teacher generated (14.1%), while 36.2% were taken from testbanks and almost half (49.4%) had no source identified. MCQs written at a lower cognitive level were significantly more likely to contain item-writing flaws. While there was no relationship between the source of the question and item-writing flaws, teachergenerated questions were more likely to be written at higher cognitive levels ($p < 0.001$). Correct answers were evenly distributed across all four options and no bias was noted in the placement of correct options. Further training in item-writing is recommended for all faculty members who are responsible for developing tests. Pre-test review and quality assessment is also recommended to reduce the occurrence of item-writing flaws and to improve the quality of test questions.

* Corresponding author. Tel.: +852 2819 2643; fax: +852 2872 6079.
E-mail address: tarrantm@hku.hk (M. Tarrant).

0260-6917/$ - see front matter
doi:10.1016/j.nedt.2006.07.006

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

Introduction

Health science disciplines, including nursing, rely heavily on multiple-choice questions (MCQs) as a method of student assessment. Most tests and examinations are developed in-house by faculty who teach the courses. Few faculty, however, have adequate education and training in developing high-quality MCQs. Teachers either develop MCQs themselves or rely on MCQ testbanks as a source of questions, both of which may result in question quality which is less than optimal. Hence, there are often substantial deficiencies in tests prepared by course teachers (Mehrens and Lehmann, 1991). Research has previously documented the poor quality of nursing testbank MCQs derived from nursing textbooks (Masters et al., 2001). Little research, however, has previously examined the quality of in-house developed nursing assessments. The purpose of this study was to examine the quality of MCQs used in teacher-developed student assessments in a baccalaureate nursing program over a five-year period from 2001 to 2005.

Review of the literature

MCQ development

MCQs consist of a question or problem statement, also referred to as the stem, and a series of four or five responses, of which only one is the correct answer. Incorrect responses are referred to as distracters and should be written as equally plausible answers to the question. The MCQ format offers many advantages for teachers. Logistically MCQs allow teachers to assess large numbers of candidates with minimal human intervention (McCoubrie, 2004). Furthermore, MCQs are objective, they allow teachers to test a wide range of content, and if well constructed, MCQs can accurately discriminate between high- and low-ability students (Schuwirth and van der Vleuten, 2003). High-quality MCQs, however, are time consuming to construct. Farley (1989a) estimates that it requires about one hour to construct a good MCQ. Other critics contend that MCQs more often test factual recall over higher cognitive thinking (Pamplett and Farnhill, 1995). Poorly constructed MCQs also frequently contain cues that allow students to guess the correct answer without the prerequisite knowledge (Downing, 2002). Hence, in reality, MCQs often poorly discriminate between high- and low-ability students.

Guidelines for MCQ construction are clearly documented in the many books and articles that have been written on this topic from a variety of disciplines (Demetrulias and McCubbin, 1982; Ellsworth et al., 1990; Gaberson, 1996;Haladyna, 2004;Haladyna and Downing, 1989a; Haladyna and Downing, 1989b; Haladyna et al., 2002; Morrison and Free, 2001; Osterlind, 1998). One of the most common problems affecting MCQs is the presence of item-writing flaws. Item-writing flaws (IWFs) are violations of these accepted item-writing guidelines which can affect student performance on MCQs, making the question either easier or more difficult to answer (Downing, 2005). Another issue affecting the quality of MCQs is that many are written at low cognitive levels. In health-science disciplines, we expect professionals to process large amounts of complex information which is used to make decisions about patient care (Masters et al., 2001). If student assessments do not test these complex cognitive functions, however, we cannot have any confidence that students will be able to perform at higher cognitive levels when required.

While much has been written in the nursing literature on developing good MCQs (Demetrulias and McCubbin, 1982; Farley, 1989b; Flynn and Reese, 1988; Gaberson, 1996; Gronlund, 1998; King, 1978; Morrison and Free, 2001), the research that has examined the actual quality of MCQs used in assessments is sparse. A search of the CINAHL database located only one study evaluating the quality of MCQs in nursing. In that study, Masters et al. (2001) examined the quality of MCQs in testbanks accompanying nursing textbooks and found 2233 item-writing flaws in 2913 nursing testbank questions. Furthermore, 72.1% of the questions were written at knowledge and comprehension levels only. In medicine, Jozefowicz et al. (2002) evaluated the quality of in-house developed examinations at three US medical schools and found that the overall quality of the questions used was low. Questions written by faculty trained in MCQ construction were of significantly higher quality than those written by untrained faculty. Downing (2005) assessed the quality of four examinations given to medical students in a US medical school and found that 46% of these MCQs contained IWFs. As a consequence, 10–15% of students who were classified as failures would have been classified as pass if items with IWFs were removed. Poor quality MCQs is a problem that not only affects nursing and health science disciplines. In an analysis of the quality of MCQs found in accounting testbanks, Hansen (1997) found that 75% of questions violated at least one item-writing guideline. Ellsworth et al. (1990) found that approximately 60% of MCQs in instructor guides accompanying introductory psychology textbooks contained IWFs.

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

Some researchers have also examined the placement of correct answers in MCQs to ascertain whether there is bias toward certain options more frequently being correct. If there is no bias we would expect that in a four option test, each option would be correct approximately 25% of the time and in a five-option test, each option would be correct 20% of the time. Clute and McGrail (1989), in a review of accounting testbank questions, found that correct responses were unevenly distributed across the five possible options, with E being the correct answer only 5% of the time. As well, Ellsworth et al. (1990), in their examination of 1022 four-option MCQs, found that option C was most frequently the correct answer (27.6%) and option A was least frequently correct (21.1%). Masters et al. (2001), however, found no significant differences in the placement of correct answers in their review of nursing testbanks.

Research aim

Tests are a key component of assessing students' knowledge. Since test grades affect students' educational outcomes and subsequent career paths, test items should be well constructed. If tests are not well constructed, assessments of student performance may be invalid. Good item construction is critical to accurate assessment. The overall aim of this study was to examine the quality of MCQs used in nursing assessments in baccalaureate nursing programs. Specifically, this study sought to evaluate the following quality indicators: (1) the frequency and nature of IWFs in MCQs used in nursing assessments; (2) the cognitive level of MCQs; (3) the primary sources of MCQs used in nursing assessments; (4) the distribution of correct answers; (5) the relationship between IWFs and cognitive level tested and (6) the relationship between question source and both the presence of IWFs and cognitive level.

Research method

Sample of questions

From late 2005 to early 2006, we retrieved all examinations and comprehensive tests that had been administered in two baccalaureate nursing programs in one nursing department over a 5-year period from 2001 to 2005. The topic areas covered in these tests included all clinical nursing courses along with health assessment, mental health nursing, community and public health, management and leadership, nursing research, nursing theory, and nursing foundations. We included only nursing courses and therefore, did not examine questions from biological science tests. The examinations and tests were used to assess students throughout a four-year Bachelor of Nursing pre-registration degree and a two-year Bachelor of Nursing post-registration degree. The majority of test papers consisted of three parts: MCQ items, short answer questions, and essay-type questions. For this analysis, we extracted the MCQs from eligible assessments, determined the source of the questions, identified duplicate questions, and examined all questions for IWFs, cognitive level tested, and the distribution of correct responses.

A total of 2770 questions were retrieved for analysis, of which 75% ($n = 2078$) were used in the pre-registration degree program and 25% ($n = 692$) were used in the post-registration degree program. Of these 2770 questions, 2174 were unique items and a further 596 questions were duplicates that had been used on more than one occasion. To qualify as a duplicate, the question had to have the exact stem and options, except for minor typographical corrections or changes. If the stem part of the MCQ had been reworded or a response option was different, the question was considered unique. Simple rewording of MCQs can remove many IWFs, therefore, some questions that contained an IWF in one usage may have been corrected in a subsequent administration of that particular item. An analysis of the study data with the duplicate questions included and excluded produced nearly identical results. As the purpose of this study was to examine the overall quality of MCQs used in nursing assessments over a set time frame, we therefore included all available MCQs regardless of whether or not they were duplicate items.

MCQ quality assessment criteria

We reviewed the literature and identified the most cited sources for MCQ construction (Gronlund, 1998; Haladyna, 2004; Haladyna and Downing, 1989a; Haladyna and Downing, 1989b; Haladyna et al., 2002; Osterlind, 1998), from which we identified 32 commonly identified item-writing guidelines. Using these guidelines, we initially conducted a preliminary review of a random sub-sample of 250 questions to determine the most commonly occurring violations of item-writing guidelines. Some violations frequently occurred, while others were less frequent, and some were rarely or never found. In total, 19 of the cited

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

item-writing violations were found in the sub-sample of 250 MCQs. These 19 guidelines were subsequently used to evaluate the quality of the 2770 MCQs (see Appendix A).

Two levels of cognition were evaluated based on Bloom's taxonomy. Although Bloom's taxonomy has never been empirically validated, it was used in this study because no other validated taxonomy exists that can be easily applied to classroom assessment (Haladyna et al., 2002). Bloom's (1956) taxonomy specifies six domains which assess incrementally higher levels of cognitive function: knowledge, comprehension, application, analysis, synthesis, and evaluation. Only the first four of these domains can be assessed using the MCQ format (Masters et al., 2001). Knowledge and comprehension are regarded as lower cognitive domains while application and analysis are regarded as higher cognitive domains. To facilitate categorization of MCQ items and to enhance inter-rater reliability, we simplified Bloom's taxonomy and classified items into two categories: K1 or K2. A K1 item assessed only recall of facts or basic comprehension and a K2 item assessed the higher cognitive domains of application and analysis.

In addition to item writing violations and cognitive domain, we also collected data on the source of the questions and the distribution of the responses. Question sources were normally specified on the test blueprint and were classified according to three categories: teacher generated, item test-bank, and no source specified. The distribution of correct responses was examined for the overall results and according to the year of study in the program.

MCQ evaluation procedures

Four reviewers evaluated each MCQ for IWFs and cognitive level. The expertise of the reviewers included content-area expertise as well as experience in developing MCQ test items. Three of the four reviewers were also trained MCQ item-writers. To ensure reliability and consistency among the four reviewers a rigorous evaluation process was undertaken. Firstly, each reviewer examined the first 700 MCQs independently. Each of the 700 items was then discussed and reviewed during a consensus panel meeting to ensure interpretation of the IWFs and cognitive level was understood similarly among all four reviewers. Next, the remaining 2070 MCQs were examined by each reviewer independently. The 250 questions used in the preliminary review were reevaluated as part of the larger sample of questions. Each reviewer's result was entered into a statistical software program and any discordant questions were identified. Discordance on either of the IWFs or the cognitive level was identified on 15% (n = 310/2770) of the MCQs and was largely related to multiple flaws in a single question. Finally, discordant MCQs were discussed during further consensus panel meetings to reach agreement on the categorization of IWFs and cognitive level.

Data analysis

Basic frequency distributions and descriptive statistics were computed for all variables. Chi-square analysis was used to determine the relationships between categorical variables such as the presence of IWFs, cognitive level tested, question source, and year of study. All data analysis was performed using Stata version 9.1 (StataCorp, 2005).

Ethical approval

Because no human subject data was collected or analyzed in this study, the Institutional Review Board of the participating institution exempted the study from the normal ethical approval process.

Results

Item writing flaws

We evaluated a total of 2770 MCQs, using the 19 item-writing guidelines described in Appendix A. Of these 2770 questions, 1280 (46.2%) contained at least one IWF and over 12% of questions containing more than one flaw (Table 1). A total of 1683 item-writing violations were found in the 2770 questions. The most frequent violations were ambiguous or unclear information in the stem (n = 208; 7.5%), negatively worded stems (n = 192; 6.9%), implausible distracters (n = 184; 6.6%), unnecessary or gratuitous information in the stem (n = 169; 6.1%), more than one or no correct answer (n = 156; 5.6%), the longest option is correct (n =

Table 1 Total number of item writing flaws in reviewed items

Number of flaws	n (%) N = 2770
None	1490 (53.8)
One	939 (33.9)
Two	290 (10.5)
Three	40 (1.4)
Four	11 (0.4)

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

135; 4.8%), logical cues in the stem ($n = 128$; 4.6%), and word repeats in the stem and correct answer ($n = 112$; 4.0%) (Table 2). Conversely, some IWFs were uncommon, including fill-in-blank question ($n = 15$; 0.5%), complex or K-type MCQs ($n = 9$; 0.3%), grammatical cues associated with sentence completions ($n = 8$; 0.3%), and convergence cues ($n = 6$; 0.2%).

Levels of cognition

The overwhelming majority of the questions ($n = 2522$; 91.1%) were written at the K1 level. Almost one-half of all MCQs written at the K1 level ($n = 1234$; 48.9%) had IWFs as compared with only 18.6% ($n = 46$) of those written at the K2 level ($p < 0.001$) (Table 3). The level of MCQs testing higher cognitive domains did increase significantly ($p < 0.001$) from lower-level courses to higher-level courses in both the Pre-registration and Post-registration programs (Table 4). While only 4.7% ($n = 35$) of questions in Year I of the Pre-registration program were written at the K2 level, this figure increased significantly to 7.5% and 14.9% for Years 3 and 4, respectively.

Table 2 Frequency of item writing flaws in multiple choice questions

Item writing flaw	n (%)[a] $n = 2770$
Ambiguous or unclear information	208 (7.5)
Negative worded stem (not, incorrect, except)	192 (6.9)
Implausible distracters	184 (6.6)
Gratuitous information in stem	169 (6.1)
More than one or no correct answer	156 (5.6)
Longest option is correct	135 (4.8)
Logical cues in stem	128 (4.6)
Word repeats in stem and correct answer	112 (4.0)
Unfocused stem	87 (3.1)
True/false question	77 (2.8)
Use of all of the above	50 (1.8)
Vague terms (sometimes, frequently)	48 (1.7)
Absolute terms (never, always)	47 (1.7)
Use of none of the above	27 (1.0)
Lost sequence in presentation of data	25 (0.9)
Fill-in-blank	15 (0.5)
Complex or K-type	9 (0.3)
Grammatical cues in sentence completion	8 (0.3)
Convergence cues	6 (0.2)

[a] Proportions do not add up to 100% as some items had no flaws and some items had more than one flaw.

Table 3 Relationship between cognitive level tested and item writing flaws

Cognitive level	Item writing flaws	
	No	Yes
K1 — Recall/ comprehension	1288 (51.1%)	1234 (48.9%)
K2 — Application analysis	202 (81.5%)	46 (18.6%)

$\chi^2 = 83.85$, $df = 1$, $p < 0.001$.

Question source

No question source was identified for just less than half of all questions ($n = 1368$; 49.4%), while 36.2% ($n = 1002$) were taken from testbanks and only 14.4% ($n = 400$) were teacher generated. Although there was no relationship between the question source and the presence of IWFs, teacher-generated questions were significantly more likely to be written at a higher cognitive level than questions from testbanks or unknown sources (Table 5).

Distribution of correct responses

For all questions, the proportion of correct answers was evenly distributed across all four options of A, B, C, or D (Fig. 1). Only in the Year 4 of the Pre-registration program did there appear to be any over- and under-representation of certain options. The result of the chi-square test, however, indicates that these differences were not statistically significant ($p = 0.15$).

Discussion

This was a descriptive study examining the quality of MCQ examinations developed in-house at one nursing department over a defined period of time. Other research does suggest, however, that the problem of suboptimal quality of MCQs extends beyond this setting and affects a large proportion of in-house assessments, both within health-science and other disciplines. Therefore, the findings from this study have a broader application and may be

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

Table 4 Relationship between cognitive level tested and year of program

Cognitive level	Pre-registration program				Post-registration program	
	Year 1	Year 2	Year 3	Year 4	Year 1	Year 2
K1 — Recall/comprehension	711 (95.3)	317 (96.4)	629 (92.5)	275 (85.1)	223 (88.1)	367 (83.6)
K2 — Application analysis	35 (4.7)	12 (3.7)	51 (7.5)	48 (14.9)	30 (11.9)	72 (16.4)

$\chi^2 = 76.06$; $df = 5$; $p < 0.001$.

Table 5 Relationship between question source and item writing flaws and cognitive level

Question source	Item writing flaws		Cognitive level	
	No	Yes	K1	K2
Teacher generated	216 (54.0%)	184 (46.0%)	330 (82.5%)	70 (17.5%)
Testbank	553 (55.2%)	449 (44.8%)	923 (92.1%)	79 (7.9%)
No source specified	721 (52.7%)	647 (47.3%)	1269 (92.8%)	99 (7.2%)
	$\chi^2 = 1.45$, $df = 2$, $p = 0.49$.		$\chi^2 = 42.19$, $df = 2$, $p < 0.001$.	

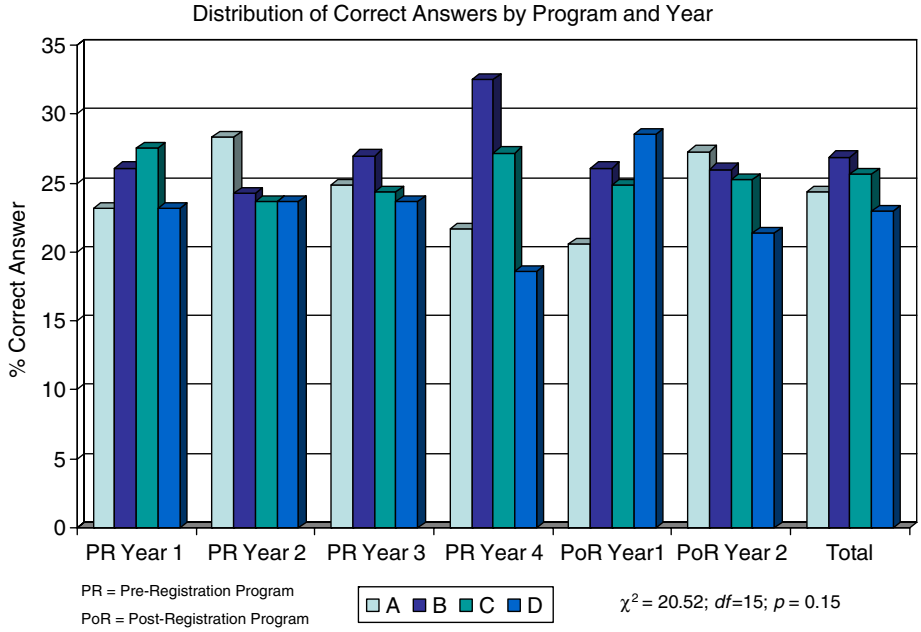

Figure 1 Proportion of correct answers for four option MCQs.

generalizable to the area of educational assessment in many disciplines.

Although high, the frequency of IWFs found in MCQs in this study (46.2%) is strikingly similar to Downing's (2005) evaluation of IWFs in medical school examination in which 46% of MCQs contained item-writing violations. The proportion of flawed questions in this study is also substantially less than the 60% found in psychology testbanks (Ellsworth et al., 1990) and the 75% found in accounting testbanks (Hansen, 1997). Unfortunately comparison of our findings with the only study conducted in nursing is not possible because Masters et al. (2001) only identified the number of individual flaws present in all of the questions and did not specify the total proportion of flawed questions. Furthermore, though almost half of all MCQs were flawed, the majority ($n = 939$) contained only one item-writing violation, suggesting that with minor editing the quality of these MCQs could be substantially improved.

Given the high proportion of questions reviewed in this study known to be from testbanks (36.2%), and the results of the studies by other researchers documenting the poor quality of testbank questions (Ellsworth et al., 1990; Hansen, 1997; Masters

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

et al., 2001), it is not surprising to find that the overall quality of MCQs was low. Testbank authors often do not have formal preparation in MCQ construction and, therefore, MCQs taken from testbanks are just as susceptible to IWFs as teacher-developed questions.

The identification of specific IWFs also highlights issues that nurse educators need to address when developing MCQ-type assessments. The high proportion of questions in this study found to be ambiguous or unclear (7.5%) may be a consequence of English as a second language (ESL) for both teachers and students. This situation is not unique to Hong Kong. Many countries where the native language is not English, nonetheless use English as the medium of instruction and assessment in nursing and health science disciplines. Teachers are required to develop valid assessments in a language that neither they nor their students use as their mother tongue, which presents a multitude of issues related to the quality and validity of those assessments. While some argue that native-language instruction may be more suitable for both teachers and students, the choice of English as the medium of instruction in most institutions is unlikely to change and needs to be examined more closely to ensure that the quality of student assessment is not compromised. Simple, clear language in both the stem and the options is highly recommended when testing ESL students as it reduces the influence of reading ability on student performance (Haladyna et al., 2002).

Other violations of item-writing guidelines that were commonly found have also been documented in the literature. Despite repeated recommendations from item-writing experts to refrain from writing negatively worded MCQs (Haladyna, 2004; Haladyna and Downing, 1989b; Haladyna et al., 2002), the use of negatives in the question stem was the second most common IWF found. Negatively worded questions are usually quicker and easier to construct, and therefore continue to be used. Research has shown, however, that this question format often performs worse than positively worded items (Haladyna and Downing, 1989b). Students can be confused by negatively worded questions, especially when they contain double negatives. Conversely, to ensure there is no ambiguity in the question, item-writers often make the correct answer (the incorrect option) so obviously incorrect that students can easily spot the answer and the question becomes too easy to adequately discriminate between the most and least able students in the test.

We also found a high proportion of questions with implausible distracters. These distracters are often added by item writers as ''fillers'' because it is difficult to come up with three equally plausible distracters in a four-option question. However, despite perceptions that MCQs must have at least four options, a question with only three plausible options is superior to a question with three plausible options and one ''filler'' option (Crehan et al., 1993; Schuwirth and van der Vleuten, 2004). Students will often detect implausible distracters and, therefore, rarely select these options.

Also of concern was that more than 5% of MCQs either had more than one correct answer or no correct answer. In single best-answer questions, it is obviously important that only one answer actually be correct. A number of violations were also found that help students correctly answer questions based on cues given in the stem or the options, rather than knowledge. IWFs such as longest option is correct, logical cues, word repeats, use of ''all of the above,'' and use of absolute terms make MCQs easier by providing helpful cues to students as to what is the correct answer. These violations favor test-wise students and potentially affect the validity of the test results.

The cognitive level of questions in this study was low with less than 10% of all questions written to test the higher cognitive domains of application and analysis. When the cognitive level of the questions was compared according to the source of questions, teacher-generated questions were significantly more likely to test a higher cognitive domain, although the proportion written at this level by teachers was still unsatisfactory. The proportion of testbank questions written at higher cognitive levels in this study (7.9%) was substantially lower than the level of 28.3% found by Masters et al. (2001). These findings should be interpreted with caution, however, as questions included in this study were neither randomly nor systematically sampled from testbanks. Testbank MCQs evaluated in this study were selected by course teachers who developed the tests and are therefore potentially subject to selection bias. In addition to MCQs, most tests extracted for this study also consisted of either short-answer questions and/or essay questions. The quality and cognitive level of these items was not examined. Therefore, if other question formats used simultaneously tested higher cognitive domains, this would help to offset the low cognitive level of the MCQ component of the overall assessment.

While many critics of MCQs argue that MCQs inherently only test regurgitation of factual material (Pamplett and Farnhill, 1995), MCQs can be written to test higher-order cognitive domains such as application and analysis (Case and Swanson,

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

2001; Schuwirth and van der Vleuten, 2003). While there is no research to suggest the optimal proportion of MCQs on a test that should be written at each level (Masters et al., 2001), it is reasonable to conclude that tests in nursing should have a high proportion of items testing higher levels of cognition as this reflects the requirements of nursing practice. While it is expected that years 3 and 4 have higher proportions of K2 items, even in lower level courses proportions of K2 questions of less than 20% seem hard to justify. It is also reasonable to expect a much higher proportion of questions used to assess post-registration students to be written at higher cognitive levels. Furthermore, while removing IWFs from MCQs does not necessarily change the cognitive domain of a question, this research suggests that writing questions at higher cognitive levels inherently removes numerous IWFs.

The trend toward discipline-based higher education in nursing means that fewer nurse academics today have had any formal preparation in educational methods such as assessment and item construction. Well-constructed MCQ items are time consuming and difficult to write. It is the responsibility of the academic institutions who hire expert clinicians to provide the necessary training and instruction to enable them to become capable faculty members (Morrison and Free, 2001). Although scant attention has been paid to item writing in the nursing literature, research in other disciplines has shown that training improves the quality of MCQs developed by teaching faculty (Hansen, 1997; Jozefowicz et al., 2002). The heavy reliance on testbanks as a source of MCQs, when these have previously been shown to be of low quality, also needs to be questioned. Furthermore, Jozefowicz et al. (2002) point out that while teachers spend considerable time planning lectures and course materials for students, insufficient time is allocated for test preparation and review prior to administration. Consequently, many tests are administered to students without adequate pre-test quality assurance. Prior to administration, a review process that includes peer review or review by an examinations committee whose members have adequate preparation in item writing, can eliminate many IWFs and ensure that a sufficient number of questions testing higher cognitive domains are included in the test. After pre-test quality assurance is complete and the test is administered, the performance of each MCQ should be examined using item analysis procedures (Farley, 1990; Jenkins and Michael, 1986). Haladyna (2004) estimates that teachers and test developers can expect that 50% of the items they write will fail to perform as expected. Therefore, item analysis will provide valuable data for question improvement and should be incorporated into the process of test development and review. To ensure valid and high-quality assessments, rigorous procedures need to be implemented to review test quality prior to administration and to review test results after administration.

Student outcomes in nursing and other educational programs are at least partially determined by the results of tests that include MCQs. If significant numbers of these questions are of low quality, then our measure of students' performance with these tests is of questionable validity. While some academics may argue that regardless of the quality of the questions, good students do well and poor students do not, this has not been empirically demonstrated to date (Jozefowicz et al., 2002). In professional programs, such as nursing, teachers are accountable to many stakeholders, including licensing bodies and the public (Crossley et al., 2002). There is therefore, an ethical and legal responsibility to ensure that assessments are of high quality and are valid.

Limitations

A limitation of this study was that was that the question source was not available for all MCQs that were evaluated, primarily exams and tests prior to 2004. Consequently, generalizations about the quality of teacher-generated versus testbank questions should be interpreted with caution. Furthermore, issues of ESL were not adequately addressed, either for teachers generating the test questions or for students taking the tests. ESL issues, however, remain vital for nursing programs using English as the medium of instruction when it is not the native language and for any non-native English language students in English language nursing programs.

Conclusion

Although this study only examined the quality of questions in one nursing department, other research suggests this problem is neither isolated to this one program nor to nursing or health science disciplines. Both the presence of IWFs and low cognitive level of MCQs are unfortunately all too common in teacher-developed examinations across many disciplines. Therefore, it is vital that teachers are provided adequate training in writing MCQ items and that all tests and examinations are subjected to adequate review both prior to and after administration.

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006

Acknowledgments

Funding for this study was provided by University of Hong Kong Run Run Shaw Research and Teaching Endowment Fund. Special thanks to Ms. Cher Lau and Ms. Winnie Lo for their assistance retrieving the MCQs for analysis.

Appendix A

Recommended guidelines for writing high-quality multiple choice questions

1. All options should be grammatically consistent with the stem and should be parallel in style and form. Non-grammatically correct options provide cues to the students who easily eliminate distracters that do not flow grammatically with the stem.
2. Each MCQ should have a clear and focused question. Teachers should avoid using MCQs with unfocused stems which do not ask a clear question or state a clear problem in the sentence completion format.
3. Each MCQ should have the problem in the stem of the question, not in the options. The options should not be a series of true/false statements.
4. The basic format for MCQs is the single best answer. Therefore, ensure that questions have one, and only one, best answer.
5. Avoid gratuitous or unnecessary information in the stem or the options. If a vignette is provided with the MCQ, it should be required to answer the question.
6. Avoid complex, or K-type MCQs. K-type MCQs have a range of correct responses and then ask students to select from a number of possible combinations of these responses. Students can often guess the answer by eliminating one incorrect response and all options containing this response or by selecting the responses which appear most frequently in all of the options.
7. Questions and all options should be written in clear, unambiguous language. Poorly worded or ambiguous questions can confuse even knowledgeable students and cause them to answer incorrectly.
8. Make all distracters plausible as plausible distracters are vital to high quality MCQs. Students who do not know the material increase their chances of guessing the correct option by eliminating implausible distracters.
9. Avoid repeating words in the stem and the correct option. Similar wording allows students to identify the correct option without knowing the material.
10. Avoid providing logical cues in the stem and the correct option that can help the student to identify the correct option without knowing the material. An example of a logical cue is asking students to select the most appropriate pharmaceutical intervention for a problem and only having one or two options which are actually pharmaceutical interventions.
11. Avoid convergence cues in options where there are different combinations of multiple components to the answer. Question writers tend to use the correct answers more frequently across all options and students will identify as correct the answer in which all components appear most frequently.
12. All options should be similar in length and amount of detail provided in the option. If one option is longer, includes more detailed information, or it contains more complex language, students can usually correctly assume that this is the correct answer.
13. Arrange MCQ options in alphabetical, chronological, or numerical order. (We assess for chronological and numerical, but not alphabetical order).
14. Options should be worded to avoid the use of absolute terms (e.g., never, always, only, all) as students are taught that there are often no absolute truths in most health science subjects and they can therefore eliminate these distracters.
15. Options should be worded to avoid the use of vague terms (e.g., frequently, occasionally, rarely, usually, commonly) as these terms lack precision and there is seldom agreement on the actual meaning of ''often'' or ''frequently''.
16. Avoid the use of negatives (e.g., not, except, incorrect) in the stem as they poorly assess students actual knowledge. If teachers wish to assess contraindications, the questions should be worded clearly to indicate that this is what is being assessed.
17. Avoid the use of ''all of the above'' as the last option. Students can easily identify if this is the correct answer by simply knowing that at least two of the options are correct. Similarly, they can eliminate it by knowing if only one of the options is incorrect.
18. Avoid the use of ''none of the above'' as the last option as it only measures students' ability to detect incorrect answers. Furthermore, if ''none of the above'' is the correct option,

the teacher must be certain that there are no exceptions to any of the options that the student may detect.

19. Avoid fill-in-the-blank format whereby a word is omitted in the middle of a sentence and the student must guess the correct word. All options should be placed at the end of the stem.

References

Bloom, B.S., 1956. Taxonomy of Educational Objectives.Handbook 1: The Cognitive Domain. Longman, London.

Case, S.M., Swanson, D.B., 2001. Constructing Written Test Questions for the Basic and Clinical Sciences. National Board of Medical Examiners, Philadelphia, PA.

Clute, R.C., McGrail, G.R., 1989. Bias in examination test banks that accompany cost accounting texts. Journal of Education for Business 64, 245–247.

Crehan, K.D., Haladyna, T.M., Brewer, B.W., 1993. Use of an inclusive option and the optimal number of options for multiple-choice items. Educational and Psychological Measurement 53 (1), 241–247.

Crossley, J., Humphris, G., Jolly, B., 2002. Assessing health professionals. Medical Education 36 (9), 800–804.

Demetrulias, D.A.M., McCubbin, L.E., 1982. Constructing test questions for higher level thinking. Nurse Educator 7 (5), 13–17.

Downing, S.M., 2002. Assessment of knowledge with written test forms. In: Norman, G.R., Van der Vleuten, C., Newble, D.I. (Eds.), International Handbook of Research in Medical Education. Kluwer Academic Publishers, Dordrecht, pp. 647–672.

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 (2), 133–143.

Ellsworth, R.A., Dunnell, P., Duell, O.K., 1990. Multiple-choice test items: what are textbook authors telling teachers? Journal of Educational Research 83 (5), 289–293.

Farley, J.K., 1989a. The multiple-choice test: developing the test blueprint. Nurse Educator 14 (5), 3–5.

Farley, J.K., 1989b. The multiple-choice test: writing the questions. Nurse Educator 14 (6), 10–12.

Farley, J.K., 1990. Item analysis. Nurse Educator 15 (1), 8–9.

Flynn, M.K., Reese, J.L., 1988. Development and evaluation of classroom tests: a practical application. Journal of Nursing Education 27 (2), 61–65.

Gaberson, K.B., 1996. Test design: putting all the pieces together. Nurse Educator 21 (4), 28–33.

Gronlund, N.E., 1998. Assessment of student achievement. Allyn and Bacon, Boston, MA.

Haladyna, T.M., 2004. Developing and Validating Multiple-choice Test Items. Lawrence Erlbaum, Mahwah, NJ.

Haladyna, T.M., Downing, S.M., 1989a. A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 2 (1), 37–50.

Haladyna, T.M., Downing, S.M., 1989b. Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 2 (1), 51–78.

Haladyna, T.M., Downing, S.M., Rodriguez, M.C., 2002. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 15 (3), 309–334.

Hansen, J.D., 1997. Quality multiple-choice test questions: Item writing guidelines and an analysis of auditing test banks. Journal of Education for Business 73 (2), 94–97.

Jenkins, H.M., Michael, M.M., 1986. Using and interpreting item analysis data. Nurse Educator 11 (1), 10–14.

Jozefowicz, R.F., Koeppen, B.M., Case, S., Galbraith, R., Swanson, D., Glew, R.H., 2002. The quality of in-house medical school examinations. Academic Medicine 77 (2), 156–161.

King, E.C., 1978. Constructing classroom achievement tests. Nurse Educator 3 (5), 30–36.

Masters, J.C., Hulsmeyer, B.S., Pike, M.E., Leichty, K., Miller, M.T., Verst, A.L., 2001. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. Journal of Nursing Education 40 (1), 25–32.

McCoubrie, P., 2004. Improving the fairness of multiple-choice questions: a literature review. Medical Teacher 26 (8), 709–712.

Mehrens, W.A., Lehmann, I.J., 1991. Measurement and Evaluation in Education and Psychology. Holt, Rinehart and Winston, Fort Worth TX.

Morrison, S., Free, K.W., 2001. Writing multiple-choice test items that promote and measure critical thinking. Journal of Nursing Education 40 (1), 17–24.

Osterlind, S.J., 1998. Constructing Test Items: Multiple-choice, Constructed-response, Performance, and Other Formats. Kluwer Academic Publishers, Boston.

Pamplett, R., Farnhill, D., 1995. Effect of anxiety on performance in multiple-choice examinations. Medical Education 29, 298–302.

Schuwirth, L.W.T., van der Vleuten, C.P.M., 2003. ABC of learning and teaching in medicine: written assessment. BMJ 326 (7390), 643–645.

Schuwirth, L.W.T., van der Vleuten, C.P.M., 2004. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education 38 (9), 974–979.

StataCorp, 2005. Stata Statistical Software: Release 9. StataCorp LP, College Station, TX.

Please cite this article as: Marie Tarrant et al., The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, Nurse Education Today (2006), doi:10.1016/j.nedt.2006.07.006