Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting

Stephen D Schneid,1 Chris Armour,2 Yoon Soo Park,3 Rachel Yudkowsky3 & Georges Bordage3

OBJECTIVES Despite significant evidence supporting the use of three-option multiple-choice questions (MCQs), these are rarely used in written examinations for health professions students. The purpose of this study was to examine the effects of reducing four- and five-option MCQs to three-option MCQs on response times, psychometric characteristics, and absolute standard setting judgements in a pharmacology examination administered to health professions students.

METHODS We administered two versions of a computerised examination containing 98 MCQs to 38 Year 2 medical students and 39 Year 3 pharmacy students. Four- and five-option MCQs were converted into three-option MCQs to create two versions of the examination. Differences in response time, item difficulty and discrimination, and reliability were evaluated. Medical and pharmacy faculty judges provided three-level Angoff (TLA) ratings for all MCQs for both versions of the examination to allow the assessment of differences in cut scores.

RESULTS Students answered three-option MCQs an average of 5 seconds faster than they answered four- and five-option MCQs (36 seconds versus 41 seconds; p = 0.008). There were no significant differences in item difficulty and discrimination, or test reliability. Overall, the cut scores generated for three-option MCQs using the TLA ratings were 8 percentage points higher (p = 0.04).

CONCLUSIONS The use of three-option MCQs in a health professions examination resulted in a time saving equivalent to the completion of 16% more MCQs per 1-hour testing period, which may increase content validity and test score reliability, and minimise construct under-representation. The higher cut scores may result in higher failure rates if an absolute standard setting method, such as the TLA method, is used. The results from this study provide a cautious indication to health professions educators that using three-option MCQs does not threaten validity and may strengthen it by allowing additional MCQs to be tested in a fixed amount of testing time with no deleterious effect on the reliability of the test scores.
Reducing options on multiple-choice questions

INTRODUCTION

Three-option multiple-choice questions (MCQs) in health professions education are still used infrequently despite a variety of evidence supporting their use. Often, MCQs contain several options that are rarely or never selected by examinees, leaving de facto only three functional options. Non-functional options are not plausible to even minimally competent students and increase the time spent on each MCQ without making any contribution to item discrimination. For newly generated MCQs, test writers intentionally avoid three-option MCQs by discarding any questions that have only three plausible options or by adding fillers such as ‘All of the above’ and ‘None of the above’. These item-writing practices are undesirable because they introduce construct-irrelevant variance into the assessment. The most likely explanation for the aversion to three-option MCQs is an exaggerated fear among educators that increased successful guessing by minimally competent students might result in the inappropriate passing of some students who should fail. A thoughtful consideration of the multiple effects of using three-option MCQs on validity evidence is necessary to help test developers make informed decisions about whether or not it is reasonable to continue to exclude these MCQs from written assessments.

One of the major arguments in favour of using MCQs with three options is that more MCQs can be included in a test for a fixed period of testing time because each three-option MCQ takes less time to complete. Proponents of the use of MCQs with fewer options argue that any potential small decrease in reliability would be offset by the increase in the number of MCQs completed. For example, Aamodt and McShane estimated that a 100-item MCQ test could be lengthened by 12% if the number of options were reduced from four to three per MCQ. They based their estimation on mean test completion times across four studies. However, they did not collect response times for individual MCQs. Owen and Froman reported that 17% more three-option MCQs could be added after comparing test response times for five- and three-option MCQs on a 100-MCQ test. In their study, undergraduate psychology students were asked to record on their answer sheets the time shown on a clock after completing the 50th and 100th MCQs. Although the time savings estimated from these two studies are similar, the methods of recording time were crude and subject to human error.

Building on previous studies, Swanson et al. examined computer-based tests. They converted 40 extended-matching questions (EMQs) used previously in the US Medical Licensing Examination (USMLE) Step 2 to a one-best-answer format. They found that MCQs with more options were more difficult than MCQs with fewer options, item discrimination did not differ significantly as the number of options decreased, and response times were longer for items with more options. They followed these findings with a similar study that examined response times when the number of options was reduced to three. They found that 55 three-option MCQs rather than 48 five-option MCQs could be administered in 1 hour, thereby allowing for a 15% increase in the number of MCQs. The three-option MCQs were about 3% easier (mean p-value of 78% versus 75%) and about 0.04 points less discriminating (mean item–total correlations of 0.21 versus 0.25). Only the difference in item difficulty was statistically significant. Differences in test score reliability were not reported.

Although the conclusions from the two studies by Swanson et al. suggest that time savings will result when fewer options are used, the method of option reduction was based on the removal of the least frequently used options using item analysis data, which does not represent a strict criterion. For example, when an option is selected by fewer than 5% of examinees, the option is considered so implausible that it is rendered non-functional. The studies also used items from national licensing examinations, which made it difficult to generalise results to locally developed medical school or health professions examinations. Furthermore, these studies like Rodriguez’s meta-analysis of eight decades of measurement research on three-option MCQs, did not address effects on pass/fail decisions.

Test scores require evidence of validity in order to be interpreted meaningfully and accurately. Specifically, the following sources of construct validity evidence, based on the current multifaceted definition of construct validity as described in the Standards for Educational and Psychological Testing, were considered as part of the conceptual framework for this study: content (representativeness of the items to domain); internal structure (item difficulty and discrimination, and test score reliability), and consequences (reasonableness of method of establishing the pass/fail line). The purpose of this study was to build on the sparse literature in health professions
education on the validity effects of using three-option MCQs created by removing non-functioning options. This research adds to the studies by Swan-son et al. by examining response times when eliminating options from previously used MCQs rather than EMQs converted into MCQs. The study addressed the following three research questions: (i) Do health professions students, in this case, Year 2 medical students (MS2s) and Year 3 pharmacy students (PS3s), respond faster to three-option pharmacology MCQs than they do to the same MCQs with the original four or five options? (ii) Do MS2s and PS3s who answer three-option pharmacology MCQs generate different data on item difficulty (p-value), item discrimination (point-biserial) and test score reliability compared with MS2s and PS3s who answer the same MCQs with the original number of options? (iii) Do medical and pharmacy faculty who judge three-option pharmacology MCQs make different judgements and establish different pass/fail cut score decisions than they do with four- or five-option MCQs?

METHODS

The institutional review boards of the University of California San Diego (UCSD) and the University of Illinois at Chicago granted ethical approval for this study.

Examination construction

Ninety-eight MCQs were used to construct two versions of an examination (Exam A and Exam B). The MCQs tested pharmacology knowledge related to six major areas (cardiology, pulmonology, gastroenterology, nephrology, endocrinology, haematology). These MCQs had been used previously in summative examinations in the PS2 core curriculum and were written by two of the authors (SDS, CA). All the MCQs were reviewed by SDS and CA to ensure they adhered to best practices of item writing. For example, all questions contained a clinical vignette and were focused on a single important topic with homogeneous options. Most importantly, item statistics were available to identify options, based on established norms, that fewer than 5% of PS2 examinees selected as the answer. A sample four-option MCQ along with its item statistics is shown in Fig. 1. For four-option MCQs that included more than one non-functional option, and for five-option MCQs that included more than two non-functional options, the options selected by the lowest proportions of students were selected for removal. In these cases, not all of the non-functional options were completely eliminated from the experimental MCQs.

Twenty-six MCQs were common to both versions of the examination and tested the six pharmacology

![Figure 1](https://example.com/figure1.png)
topics in proportion to their representation in the entire examination. These common MCQs were used to determine the equivalence of the two randomly assigned groups of students (taking either Exam A or Exam B) in terms of performance and test-taking speed.

The remaining 72 MCQs were divided into two experimental sets of 36 MCQs (Set 1 and Set 2) by randomly dividing the items on each topic into two categories of long and short items. The 36 MCQs in Set 1 contained the original four or five options (‘long’ MCQs) in Exam A, and were converted to three-option questions (‘short’ MCQs) by the removal of non-functional options for Exam B. The other 36 MCQs in Set 2 contained the long MCQs in Exam B and were converted to short MCQs for Exam A. Table 1 shows these details.

Thus, half of the ‘experimental MCQs’ from each pharmacology topic were selected to have the original number of options and the other half to have three options. Approximately 50 options were removed in the process of creating the three-option items for Experimental Set 1 in Exam A and Experimental Set 2 in Exam B. In Exam A, Experimental Set 2 had 22 four-option MCQs and 14 five-option MCQs. In Exam B, Experimental Set 1 had 23 four-option MCQs and 13 five-option MCQs. Exams A and B were administered using the Daskala Software Platform (Daskala, LLC, Chicago, IL, USA), which provides item response times and item analysis data.

Student participants

In the spring of 2012, MS2s (n = 125) and PS3s (n = 60) at UCSD were solicited, via a class announcement, to take a 98-MCQ computerised pharmacology examination. A total of 39 PS3s and 38 MS2s participated in the study, resulting in an overall participation rate of 42% (77 of 185 students). All the students had learned pharmacology in previous courses prior to the study and were given a list of relevant learning objectives and drugs several weeks before the examination to help them prepare. From the students’ perspective, the only purpose of the examination was self-assessment. The students provided consent upon participation and were able to withdraw at any stage.

Examination administration

The students were randomly assigned to take either Exam A or Exam B. The examinations were administered using a ‘closed book’ protocol on secure computers provided by the UCSD Biomedical Library computer laboratory. Multiple proctored testing periods were offered over a 2-week period; participants agreed not to discuss the examination with their peers. The students were allowed 150 minutes (average of 1.5 minutes/MCQ) to answer the 98 MCQs, which was a sufficient amount of time to finish the examination without time pressure. The students were instructed to answer each question, to move forward, and not to return to any of the questions. All students who volunteered to participate completed the examination.

Student performance data analysis

Each version of the examination was divided into three MCQ sets (Control Set, Experimental Set 1, Experimental Set 2; see Table 1). For each MCQ set, four indices were calculated: mean response time in seconds; mean item difficulty, calculated as the percentage of MS2s and PS3s who responded correctly to the item; mean item discrimination, calculated as the point-biserial index (or correlation between student performance on the MCQ [i.e., 1 = correct, 0 = incorrect] and performance on the entire examination), and test score reliability according to the Kuder–Richardson formula 20 (KR 20). Data on performance on short and long MCQs from Exams A and B were pooled to generate a 72-MCQ short MCQ examination and a 72-MCQ long MCQ examination because performance on the items common to the two examinations was similar across both examinations. Unpaired two-tailed t-tests were used to compare the differences in mean response time, item difficulty and discrimination. We also assessed how many of the removed
options remained non-functional for each experimental set of MCQs.

Faculty participants

For the standard setting part of the study, UCSD health science faculty members, with expertise in the subject matter, were recruited to judge each MCQ in Exam A and Exam B. Five pharmacy faculty judges (all with a PharmD) and six medical faculty judges (five with MDs and one with a PhD) participated in the study. None of the judges had been involved in the construction of the MCQs.

Standard setting

The medical and pharmacy faculty judges separately participated in two standard setting sessions: in one standards were set for Exam A, and in the other standards were set for Exam B. A washout period of 9–10 weeks was observed between the two sessions. As the judges reviewed all MCQs in each examination, each judge served as his or her own control. The 26 MCQs common to both versions of the examination were used to assess the reproducibility of the judgements after the washout period.

Each standard setting session involved a 35-minute orientation and training component that included a discussion of the stakes involved and the definition of a minimally competent PS3 or MS2, and some practice in making judgements with MCQs. Pharmacy faculty members made judgements based on their agreed-upon description of a minimally competent PS3, and medical faculty staff made judgements based on their agreed-upon description of a minimally competent MS2.

Judges were given a paper copy of the examination in which the item difficulty was listed next to each MCQ and were asked to indicate whether a minimally competent student would answer the MCQ correctly using a three-level Angoff procedure by indicating ‘Yes’, ‘No’ or ‘50–50’. ‘Yes’ judgements were assigned a value of 1; ‘No’ judgements were assigned a value of 0, and ‘50–50’ judgements were assigned a value of 0.5. Mean judgements were calculated for each MCQ and summed to generate cut scores for the different examination subsets. The three-level Angoff procedure was chosen because it provides more sensitivity to changes in judgements than the ‘Yes/No’ Angoff method and is less cognitively challenging for judges than making multiple levels of judgements (e.g. 0–100%).

Standard setting data analysis

Cut scores were calculated for each MCQ set separately using the medical and pharmacy faculty judgements. All judgements on short MCQs from Exams A and B were combined to generate a 72-short MCQ health professions cut score. All judgements on long MCQs from Exams A and B were combined to generate a 72-long MCQ health professions cut score. An unpaired two-tailed t-test was used to compare the mean cut scores for control and experimental sets of questions.

RESULTS

Common MCQs

The mean response times, item difficulty and item discrimination indices did not differ for the 26 items common to the two versions of the examination (p-values > 0.46). There was a 0.07 difference in test score reliabilities (KR 20: 0.70 versus 0.63). Table 2 shows details. Consequently and under the assumption that the students taking Exams A and B were equivalent, the data for the experimental sets in Exams A and B were combined for a total of 72 short and 72 long MCQs.

Short MCQs versus long MCQs

Response time

Overall, it took students a mean ± standard deviation (SD) of 36 ± 12 seconds to answer a short MCQ and 41 ± 12 seconds to answer a long MCQ; the 5 seconds per MCQ difference was statistically significant (p = 0.01; effect size 0.45). The effect was greater when five options were converted to three options; for example, students took an average of 8 seconds less per MCQ for the 27 three-option MCQs compared with the five-option MCQs (p = 0.004).

Item characteristics

Mean item difficulty (percentage correct) did not differ statistically between the short (67%) and long (62%) MCQs (p = 0.11). Similarly, mean item discrimination did not differ statistically between the short (0.21) and long (0.26) MCQs (p = 0.11). Table 3 shows details.
Test score reliability

The mean test score reliability for the short MCQ examination was lower than that for the long MCQ examination (0.69 versus 0.72); the averages of the KR 20s were computed under the assumption that the data were normally distributed.

Overall, about half of the 49 and 50 options removed from Experimental Sets 1 and 2 four- and five-option MCQs (49% and 54%, respectively) were found to be non-functional.

Standard setting

For the control item set, the health professions judges generated cut scores of 49% for Exam A and 44% for Exam B. Although the 5% difference was not statistically significant (p = 0.52), this higher cut score produced a significantly higher fail rate (18.0% versus 5.3%). The overall cut score for the combined medical and pharmacy judgements was 8% higher for the short MCQs than the long MCQs (57% versus 49%; p = 0.04).

DISCUSSION

In a manner consistent with the findings of previous research, the results from this study showed that students took less time to answer three-option MCQs than to answer four- or five-option MCQs, with an overall saving of 5 seconds per MCQ. This time difference would allow for the completion of approximately 14 more questions (104 versus 90) per hour of testing time, assuming an average of 40 seconds per MCQ as found in this study. This 16% increase in MCQs is similar to the increases identified in pre-
vious studies.4,5,7 The ability to add more MCQs when testing time is limited is a major benefit of using three-option MCQs because it increases sampling and improves content validity and minimises construct under-representation.

Adding more MCQs is likely to improve test score reliability. For example, if we consider an original short MCQ test score reliability of 0.69 (as in the present study), using the Spearman–Brown formula and adding 16% more short MCQs would yield an estimated test score reliability of 0.72. The additional MCQs that could be added according to these time savings could increase reliability to that of the long MCQ examination in the present study (i.e. 0.72) and enhance content-related validity evidence. Although reducing the number of options may increase the probability that a student can randomly guess the correct answer, the three-option MCQs in this study were not significantly easier than the MCQs with more options. The reason the non-functional option elimination method was used in this study to reduce the number of items was to minimise the effects on item difficulty.

This is the first study to have looked at the effects of option reduction in MCQs on standard setting judgements and the results indicate that the judges raised the cut score for the three-option MCQs when using a three-level Angoff method. Educators should be aware that reducing the number of options may result in an increased failure rate if a test-based absolute standard setting method is used, depending on the distribution of student scores.

Although we did not debrief the judges to explore the differences they perceived between the two versions of the examination, the increased cut score may have resulted from the judges’ bias that having fewer options in the three-option MCQs would cause a significant boost in performance. This traditional perception among educators may explain why there has been continued reluctance to use three-option MCQs.

This study has some limitations. Firstly, the study may have been insufficiently powered with regard to the non-statistically significant findings. The low participation rate and number of MCQs limited our statistical interpretation of the performance data. If we had recruited 150 subjects, a number sufficient to generate stable item characteristics, approximately 150 MCQs would have been required to make the difference of about 5% in item difficulty meaningful. Secondly, although performance on the ‘control’ items was equivalent, there were no external baseline data, such as information on previous academic performance, with which to better compare the two groups taking the different versions of the examination. Students also volunteered to take the examination, which may have created some selection bias. Thirdly, as the examination was formative, rather than being high-stakes, it is difficult to generalise our findings to summative conditions. Fourthly, as all the items tested pharmacology, we cannot generalise the results to other subject areas. Finally, only about 50% of the non-functional options identified from previous item analysis taken from PS2s turned out to be non-functional for this experiment, which may have exaggerated the differences in the standard setting results.

Although there is nearly a century of research supporting the use of three-option MCQs,1 it has not had much impact on the strong orthodoxies that exist regarding the number of options used in health professions’ MCQ examinations. Unfortunately, these traditions have been remarkably resistant to change over this period. The findings from this study challenge yet again the traditional approach to standardising the number of options for an entire examination to four or five. The results replicate decades of research supporting the use of three-option MCQs, and add to the existing literature by examining the effects of option reduction on response time and standard setting judgements. Thus, this study provides a cautious indication to health professions educators that using three-option MCQs does not threaten validity and may strengthen it by allowing additional MCQs to be tested in a fixed amount of testing time with no deleterious effect on the reliability of the test scores. From a practical perspective, MCQ writers can continue to write as many plausible options as feasible, but, most importantly, they should not discard three-option MCQs during the test development phase. Therefore, an examination adhering to the recommendations of this study may contain a blend of three-, four- and five-option MCQs. The method of option reduction should involve the use of item analysis reports or content expertise to eliminate non-functional options. The practice of eliminating poor distractors, which may allow the inclusion of more items per unit of testing time, may, depending on the content of the additional test material, provide for more valid and reliable test scores.
Contributors: SDS originated the concept and design of the project, contributed to the acquisition and analysis of data, and drafted the article. CA contributed to the design of the project and analysis of data. YSP contributed to the design of the project and analysis of data, with specific reference to some of the statistics used. RY contributed to the design of the project, particularly to the standard setting arm, and the analysis of data. GB contributed to the design of the project and analysis of data. All authors contributed to the critical revision of the article and approved the final manuscript for publication. All authors are accountable for all aspects of the work.

Acknowledgements: none.

Funding: none.

Conflicts of interest: not applicable.

Ethical approval: this study was certified as exempt from full review by the institutional review boards of the University of California San Diego and University of Illinois at Chicago.

REFERENCES

1 Rodriguez M. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Meas Issues Pract* 2005;24(2):3–13.

2 Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Meas* 1993;53:999–1010.

3 Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct under-representation. *Adv Health Sci Educ Theory Pract* 2002;7:235–41.

4 Aamodt MG, McShane T. A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Pers Manage* 1992;121:151–60.

5 Owen SV, Froman RD. What’s wrong with three-option multiple-choice items? *Educ Psychol Meas* 1987;47:513–22.

6 Swanson DB, Holtzman KZ, Clauser BE, Sawhill AJ. Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Acad Med* 2005;80(Suppl):95–6.

7 Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extending-matching questions in relation to number and source of options. *Acad Med* 2008;83(Suppl):21–4.

8 Downing SM, Yudkowsky R, eds. *Assessment in Health Professions Education*. New York, NY; London: Routledge 2009:21–55.

9 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: AERA 1999.

10 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;37:830–7.

11 Haladyna TM, Downing SM. The validity of a taxonomy of multiple-choice item writing rules. *Appl Meas Educ* 1989;2(1):51–78.

12 Cizek GJ, O’Day DM. Further investigations of non-functioning options in multiple-choice test items. *Educ Psychol Meas* 1994;54(4):861–72.

13 Yudkowsky R, Downing SM, Popescu M. Setting standards for performance tests: a pilot study of a three-level Angoff method. *Acad Med* 2008;83(10)(Suppl):13–6.

14 Impara JC, Plake BS. Standard setting: an alternative approach. *J Educ Meas* 1997;34(4):353–66.

Received 1 January 2014; editorial comments to author 24 February 2014, 15 May 2014; accepted for publication 19 May 2014