Reducing the Number of Options on Multiple-Choice Questions: Response Time, Psychometrics,
and Standard Setting

BY

Stephen D. Schneid
B.S., University of California, San Diego, 2000

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Health Professions Education
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Thesis Committee:

Georges Bordage, Chair, Department of Medical Education
Yoon Soo Park
Rachel Yudkowsky
Chris Armour, University of California, San Diego

TABLE OF CONTENTS

<u>CHAPTER</u>	<u>PAGE</u>
I. Purpose	1-2
II. Methods	2
A. Exam Construction.....	2-3
B. Student Participants.....	3
C. Exam Administration.....	3
D. Student Performance Data Analysis.....	3-4
E. Faculty Participants.....	4
F. Standard Setting.....	4
G. Standard Setting Data Analysis.....	4
III. Results.....	5
A. Common MCQs.....	5
B. Short MCQs versus Long MCQs.....	5
C. Nonfunctional Options.....	5
D. Standard Setting.....	5
IV. Discussion and Conclusion.....	6-7

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I.	Mean Response Time, Item Difficulty, Discrimination, and Test Reliability for Control Questions within Exam A and Exam B.....	9
II.	Mean Response Time, Item Difficulty, Discrimination, and Test Reliability for the Three-Option and Four- and Five Option MCQ Exams.....	10
III.	Mean Cut Scores for Three-Option versus Four- and Five-Option MCQs.....	12

LIST OF BOXES

<u>BOX</u>	<u>PAGE</u>
I. Item Distribution on the Two Forms (A and B) of the MCQ Exams...	8
II. Number of Four- and Five-Option MCQs, with Percentage of Nonfunctional Options for the Two Experimental Sets.....	11

SUMMARY

A study examining the effects of reducing the number of multiple-choice questions (MCQs) on response time, psychometrics, and absolute standard setting was conducted using health professions students and faculty as research subjects. Two forms of a computerized MCQ exam was administered to second-year medical students and third-year pharmacy students. Four- and five-option MCQs were converted into three-option MCQs to create the two exam forms by removing nonfunctional options. Medical and pharmacy faculty judges provided three-level Angoff ratings for all MCQs for both exam forms to assess differences in cut scores.

Students (n=77) answered three-option MCQs, on average, 5.24 seconds faster than the four- and five-option MCQs. No significant differences in item difficulty and discrimination, or test reliability were observed. Overall, the cut scores generated for three-option MCQs were 8.32 % higher.

Test developers should be encouraged to use three-option MCQs if there are only two plausible “distractors”. Item analysis should be used to eliminate or replace nonfunctional options. Health professions educators should be aware that the cut score, and consequently the fail rate, may increase with the use of 3-option MCQs if an absolute standard setting method is used such as the three-level Angoff.

Three-option multiple-choice questions (MCQs) in health professions education are still used infrequently despite a variety of evidence supporting their use.¹ MCQs often contain several options that are rarely or never selected by examinees, leaving de facto only three functional options.² The nonfunctional options are not plausible to even the minimally competent students and only increase time spent per MCQ, without 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.” This item-writing practice is undesirable because it introduces construct-irrelevant variance (CIV) 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 would result in incorrectly passing 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 an informed decision about whether or not it is reasonable to continue to exclude these MCQs from their written assessments.

One of the major arguments in favor of using MCQs with three options is that more MCQs can be included on a test for a fixed period of testing time. Proponents of using fewer options argue that any potential small decrease in reliability would be offset by the increase in the number of MCQs tested. For example, Aamodt and McShane estimated that a 100 four-option MCQ test could be lengthened by 12% if the number of options were reduced to three per MCQ.³ 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.⁴ In their study, undergraduate psychology students were asked to record the time on their answer sheets from a clock after completing the 50th and 100th MCQ. While the time savings estimated from these two experiments is similar, the methods of recording time were crude and subject to human error.

Building on previous studies, Swanson et al examined computer-based tests.⁵ They converted 40 extended-matching questions (EMQs) used previously on USMLE Step-2 to one-best-answer format. They found that MCQs with more options were more difficult than MCQs with fewer options, and item discrimination was not significantly different as the number of options decreased. Response times were longer for items with more options. Swanson followed up with a similar study that examined response times for option reductions down to three options.⁶ They found that 55 three-option MCQs could be administered in one hour compared to 48 five-option MCQs (a 14.6% increase). Compared to the five-option MCQs, the three-option MCQs were about three percent easier (mean p-value of 77.9% versus 75.3%) and about 0.04 less discriminating (mean item-total correlation of 0.213 versus 0.246). Only the difference in item difficulty was statistically significant. Differences in test reliability coefficients were not reported.

While the results from two Swanson studies suggest that there would be time savings 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, not a strict criterion of less than a

5% response.² The studies also involved items used on national licensing exams, making it difficult to generalize results to locally-developed medical school or health professions examinations. Furthermore, these studies, as well as Rodriguez's meta-analysis of eight decades of measurement research on three-option MCQs, did not address the effects on pass/fail decisions.¹

Test scores require evidence of validity in order to be interpreted meaningfully and accurately.¹ Specifically, the following sources of construct validity evidence, based on the current multi-faceted definition of construct validity, as described in the *Standards of Educational and Psychological Testing*,^{7,8,9} were considered: 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 upon the sparse literature in health professions education regarding the validity effects of using three-option MCQs created by removing nonfunctioning options. Furthermore, this research adds to the Swanson studies 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: (1) Do health professions students, in this case, second-year medical students (MS2s) and third-year pharmacy students (PS3s), respond faster to three-option pharmacology MCQs compared to the same MCQs with the original four or five options? (2) Do MS2 and PS3 students who answer three-option pharmacology MCQs produce different item difficulty (p-value), item discrimination (d-value), and test score reliability compared to MS2 and PS3 students who answer the same MCQs with the original number of options? And, (3) do medical and pharmacy faculty who judge three-option pharmacology MCQs make different judgments and establish different pass-fail cut score decisions compared to four or five-option MCQs?

Methods

Exam Construction

Ninety-eight MCQs were used to construct two exam forms (Exam A and Exam B). The MCQs covered six major topics in pharmacology (Cardiology, Pulmonology, Gastroenterology, Nephrology, Endocrinology, and Hematology) and were previously used on summative exams in the PS2 core curriculum. These MCQs adhered to best practices of item writing,¹⁰ and item statistics were available.

Twenty-six MCQs were identical for both exams and tested the six pharmacology topics in proportion to their representation on the overall exam. Most of these questions had fewer than two nonfunctional options. These common MCQs were used to determine the equivalence of the two randomly assigned groups of students (taking either Exam A or 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 groups. 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 nonfunctional options for Exam B as described below. The other 36 MCQs in Set 2 contained the long MCQs in Exam B and were converted to short MCQs for Exam A. See Box 1 for details.

A nonfunctional option was identified through item analysis and was defined, based on established norms,^{2,11} as an option that fewer than 5% percent of examinees selected as the answer. Test statistics generated from past PS2s was used to identify nonfunctional options for removal. Half of the "experimental MCQs" from each pharmacology topic were selected to have the original number of options, 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. On Exam A, Experimental Set #2 had 22 four-option MCQs and 14 five-option MCQs. On Exam B, Experimental Set #1 had 23 four-option MCQs and 13 five-option MCQs. Exams A and B were entered into a computer-based testing program provided by Daskala©. The software provided item response times and item analysis data.

Student Participants

MS2s (n=125) and PS3s (n=60) from UCSD were solicited in March 2012, via a class announcement, to take a 98-MCQ computerized pharmacology exam. The students were given a list of relevant learning objectives and list of drugs several weeks before the exam to help them prepare. They provided consent upon participation. The institutional review boards from the University of California, San Diego (UCSD) and University of Illinois, Chicago (UIC) approved this study.

Exam Administration

The students were randomly assigned to take either Exam A or Exam B. The exams were administered on secure computers provided by the UCSD Biomedical Library (BML) computer laboratory. Multiple testing periods were offered over a two-week period; participants agreed not to discuss the exam with their peers. The students had 150 minutes (average of 1.5 minutes/MCQ) to complete the exam, a non-speeded test.

Student Performance Data Analysis

Each form of the exam was divided into three MCQ sets (Control Set, Experimental Set #1, and Experimental Set #2; see Box 1. For each MCQ set four indices were calculated: mean response time in seconds per MCQ, 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 (that is, 1 = correct and 0 = incorrect) and performance on the entire exam), and test score reliability using Kuder-Richardson Formula 20 (KR 20). Performance data from short MCQs and long MCQs from Exams A and B could later be pooled together to generate a 72 MCQ short-MCQ exam and 72 MCQ long-MCQ exam if performance on the common items

between the two exams were similar. 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 were still nonfunctional for each experimental set of MCQs.

Faculty Participants

For the standard setting part of the study, UCSD health science faculty were recruited purposefully to judge each MCQ from Exam A and Exam B. Five pharmacy faculty judges (all with a PharmD) and six medical faculty judges (five MDs and one PhD) participated in the study. The selection criterion was that they had good command of the subject matter.

Standard Setting

We held two standard setting sessions, separate for the medical and pharmacy faculty judges, one to set standards for Exam A and the other to set standards for Exam B. To reduce the potential for a recency effect, a washout period of nine to ten weeks was used between the two sessions. Since the judges reviewed all MCQs in each exam, each judge served as his/her own control. The 26 common MCQs on each exam version were used to assess the reproducibility of the judgments over the washout period.

Each standard setting session had a 35-minute orientation and training component that included a discussion of the stakes involved and the definition of a minimally competent student, and some practice. Pharmacy faculty made judgments based on their agreed-upon description of a minimally competent PS3, while the medical faculty made judgments based on their agreed-upon description of a minimally competent MS2.

Judges were given a paper copy of the exam that included the item difficulty 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.¹² They had the choices of “Yes,” “No,” and “50-50” with respect to a minimally competent student’s chance of answering the MCQ correct. “Yes” judgments were assigned a value of 1, “No” judgments were assigned a value of 0, and “50-50” judgments were assigned a value of 0.5. The mean judgments were calculated for each MCQ and summed to generate cut scores for the different exam subsets. The three-level Angoff was chosen because it provides more sensitivity to changes in judgments than the “Yes-No” Angoff method and is less cognitively challenging for the judges than multiple levels of judgments (e.g., 0-100%).

Standard Setting Data Analysis

Cut scores were calculated for each MCQ set separately using the medical and pharmacy faculty judgments. All short-MCQ judgments from Exams A and B were also combined together to generate a 72 short-MCQ health professions cut score. All long-MCQ judgments were combined together from Exams A and B to generate a 72 long-MCQ health professions cut score. A two-tailed t-test was used to compare the mean cut scores.

Results

A total of 77 health professions students (39 PS3s out of a 60-student class and 38 MS2s out of a 125-student class) participated in the study during the spring of 2012, for a participation rate of 42%.

Common MCQs

The mean response times and item difficulty and discrimination indexes did not differ for the participants completing the 26 common items on the two versions of the exam (p-values greater than .46). There was a .07 difference in test score reliabilities (KR20 = 0.70 vs. 0.63). See Table 1 for details. Consequently and under the assumption that the students taking Exam A and B were equivalent, the data for the experimental sets in Exam 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 an average of 36.11 seconds to answer a short MCQ and 41.35 second for long MCQs; the 5.24 seconds per MCQ difference was statistically significant ($p=0.008$; $ES=0.45$). The effect was greater when converting five options to three options; for example, students took 8.29 seconds less per MCQ for the 27 three-option MCQs compared to the five-option ones ($p=0.004$).

Item characteristics. Although the mean percent correct for the short MCQs was higher compared to long MCQs (67.46% and 62.47%, respectively) and the mean item discrimination for the short MCQs was lower than for the long MCQs (0.21 and 0.26, respectively), both of these differences were not statistically significant (p-values >0.105). See Table 2 for details.

Test score reliability. The mean test score reliability for the short MCQ exam was lower compared to the long MCQ exam (0.685 versus 0.720); the averages of the KR20s were computed under the assumption that the data was normally distributed.

Nonfunctional Options

Overall, about half of the options removed from experimental set #1 and set #2 four and five-option MCQs (49% and 54% respectively) were found to be non-functional; See Box 2 for details.

Standard Setting

The overall cut score for the combined medical and pharmacy judgments was 8.32% higher for the short MCQs than for the long MCQs (57.32% vs. 49.00%, $p=0.037$); see Table 3.

Discussion and conclusion

Consistent with previous research, the results from this study showed that students took less time to answer three-option MCQs compared to four or five options. With an overall saving of 5.24 seconds per MCQ. This difference would allow approximately 14 more questions (104 vs. 90) per hour of testing time, assuming an average of 40 seconds per MCQ. This 16% increase in MCQs is similar to results from previous studies mentioned in the introduction. 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 adds to content validity, by minimizing construct-underrepresentation.

Adding more equivalent MCQs will likely improve test score reliability. For example, if we consider an original short MCQ test score reliability of 0.685 (as in the present study) and using the Spearman-Brown formula, adding 16% more short MCQs would yield an estimated test score reliability of 0.716. The additional MCQs that could be added from time savings could increase reliability to the same level as for the long MCQ level in the present study (e.g., 0.72) and enhance validity evidence related to internal structure.

Even though the three-option MCQs in this study were not easier, the results from the standard setting portion of the study indicate that judges raised the cut score for the three-option MCQs when using the three-level Angoff method. While the increase in cut score followed a non-statistically significant increase in student performance, educators should be aware that reducing options would likely result in an increased failure rate using a test-based absolute standard setting method, depending on the distribution of student scores. The increased cut score is likely a result of the judges' overestimation in the boost in performance that students may get with three-option MCQs, which is consistent with faculty's hesitancy regarding the use of three-option MCQs.

There are some limitations to the study. First, since all the items tested pharmacology, one cannot generalize the results to other subject areas. Second, since all of the MCQs adhered closely to the NBME guidelines for MCQ writing, the results are difficult to generalize to the types of MCQs commonly used in the classroom setting. On the other hand, there may be even larger response time effects seen when MCQs contain certain common MCQ flaws (e.g. unfocused stem with heterogeneous options). Third, the students volunteered to take the exam, which may have created some selection bias, but performance of the students taking the two versions of the exams were equivalent. Furthermore, since the exam was formative (not high stakes), it is difficult to generalize to summative conditions. Fourth, only about 50% of the nonfunctional options identified from previous item analysis turned out to be nonfunctional for this experiment, which could have exaggerated the standard setting results. Finally, the small number of subjects and MCQs also limited our statistical interpretation of the performance data. Assuming we recruited 150 subjects, a number sufficient to generate stable item characteristics, we needed to have approximately 150 MCQs to make the difference of about 5% in the item difficulty meaningful.

Overall, the results from this study indicate that using three-option MCQs does not threaten validity and likely strengthens it by allowing additional MCQs to be tested in a fixed amount of testing time with no deleterious effect on the reliability of the scores. From a practical perspective, MCQ writers can continue to write as many plausible options as feasible and not discard three-option MCQs. Also, test developers can evaluate item analysis reports and eliminate nonfunctional options to make assessment more time efficient.

Box 1

Item Distribution on the Two Forms (A and B) of the MCQ Exams.

Table 1. Mean Response Time, Item Difficulty, Discrimination, and Test Reliability for Control Questions within Exam A and Exam B.

MCQ Set	Exam Form	Mean (SD) response time in seconds/MCQ	Mean (SD) Item Difficulty (% correct)	Mean (SD) Item Discrimination (pt. biserial)	Test score reliability (KR 20)
Control Set (26 MCQs)	Exam A	39.46 (13.71)	58.88% (21.48%)	0.20 (0.16)	0.70
	Exam B	42.00 (15.40)	61.44% (22.25%)	0.16 (0.19)	0.63
	Absolute Difference	2.54	2.56%	0.04	0.07
		p=0.54 [†]	p=0.68	p=0.459	

[†] Statistical tests done using two-tailed t-tests.

Table 2. Mean Response Time, Item Difficulty and Discrimination for the Three-option and Four- and Five-Option MCQ Exams.

	Mean (SD) response time in seconds/MCQ	Mean (SD) Item Difficulty (% correct)	Mean (SD) Item Discrimination (point- biserial)	Test Score Reliability (KR 20)
3-option exam (72 MCQs)	36.11 (11.51)	67.46% (18.01%)	0.21 (0.155)	0.720
4-and 5-option exam (72 MCQs)	41.35 (11.86)	62.47% (19.16%)	0.26 (0.185)	0.685
Absolute Difference	5.24	4.99%	0.05	0.035
	p=0.008 [†]	p=0.11	p=0.105	

[†] Statistical tests done using two-tailed t-tests.

Box 2. Number of Four- and Five-Option MCQs, with Percentage of Nonfunctional Options for the Two Experimental Sets.

Experimental Set #1

23 four-option MCQs (8 nonfunctional options/23 options removed)

13 five-option MCQs (16 nonfunctional options/26 options removed)

24 non-functional options/49 options removed = 49%

Experimental Set #2

22 four-options MCQs (12 nonfunctional options/22 options removed)

14 five-option MCQs (15 nonfunctional options/28 options removed)

27 non-functional options/50 options removed = 54%

Table 3. **Mean Cut Scores for Three-Option versus Four- and Five-option MCQs.**

Item Set	Exam Form	Cut score
		Mean (SD)
Experimental Set (72 MCQs)	3-option Exams A and B	57.32% (7.41%)
	4- and 5-option Exams A and B	49.00% (5.00%)
	Difference	8.32% p=0.037 [†]

[†] Statistical tests done using two-tailed t-tests.

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; 3-13.
2. Haladyna TM, Downing SM. How Many Options is Enough for a Multiple-Choice Test Item? *Educational and Psychological Measurement.* 1993; 53: 999-1010.
3. Aamodt, MG, McShane T. A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management.* 1992;121:151-60.
4. Owen SV, Froman RD. What's wrong with three-option multiple-choice items? *Educational and Psychological Measurement.* 1987; 47: 513-522.
5. 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: S93-S96.
6. 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: S21-S24.
7. Downing, S.M. & Yudkowsky, R. (Eds) *Assessment in Health Professions Education.* New York & London: Routledge; 2009.
8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association; 1999.
9. Downing, S. M. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003; 37: 830-837.
10. Haladyna TM, Downing SM. (1989b), Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education* 2: 51-78
11. Cizek GJ, O'Day DM. Further investigations of nonfunctioning options in multiple-choice test items. *Educ Psychol Meas.* 1994; 54(4): 861-72.
12. Yudkowsky, Rachel; Downing, Steven M.; Popescu, Mihaela. Setting standards for performance tests; a pilot study of a three-level Angoff method. *Acad Med.* 2008; 83(10):S13-S16.

Stephen D. Schneid

6165 Caminito Baeza, San Diego, CA 92122
Home Phone: (858) 354-9262 Work Phone (858) 534-6962
e-mail: sschneid@ucsd.edu,
http://meded.ucsd.edu/ugme/oess/contact_information/

Education

University of California, San Diego
Animal Physiology/Neuroscience, B.S., and Biopsychology, B.S.
Graduated Cum Laude June, 2000

Teaching Experience

Physician Retraining and Reentry, Inc
Oct/'12-present

Director of Assessment

UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences (SSPPS)
Sept/'10-present

Co-director of the Team-Based Learning (TBL) program

Co-director and Instructor for SSPPS 239 (Pharmacy Topics in Endocrinology)

UCSD School of Medicine (SOM) Office Educational Support Services (OESS)
June/'10-present

Educational Consultant for the Primary Care Clerkship

Sept/'02-present

Health Sciences Educator

June/'02-June/'10

Lecturer for the United States Medical License Examination (USMLE) STEP 1 Review Course

June/'01-present

Lecturer for Summer Prematriculation Course for medical and pharmacy students

Teaching Accomplishments

SSPPS

Instructor of the year (voted by second-year pharmacy students), **2009-10**

UCSD School of Medicine

Kaiser Teaching Award nominee, 2011-12

Vice Chancellor's UCSD Health Sciences Employee of the Year Honoree, 2007

Vice Chancellor's UCSD Health Sciences Employee of the Year Honoree, 2004

UCSD Department of Biology

Paul Saltman Teaching Award, 2000

Educational Scholarship

Publications

Schneid SD; Kingston PA; Apperson A. Simplified Bookmark Method for Local Medical School Exams. *Medical Education*. 2011; 45(5); 533-34.

Posters and presentations

Schneid SD; Kingston PA; Apperson A; Armour JC; Laiken N, "Development of a shared, integrated assessment bank for an interprofessional health sciences curriculum", Regional Conference of the Western Group on Education Affairs (2010)

Schneid SD; Kingston PA; Apperson A; Armour JC; Laiken N, "Implementation of interprofessional team-based Learning (TBL) in the pre-clerkship curriculum", Annual Meeting of the Association of American Medical Colleges (2010)

Schneid SD; Armour JC, "Implementation of modified team-based learning (TBL) to support pharmacy students in a shared curriculum", Regional Conference of the Western Group on Education Affairs (2011)

Schneid SD; Kingston PA; Apperson A, "Simplified Bookmark Method for Local Medical School Exams", Regional Conference of the Western Group on Education Affairs (2011)