Commentary

Journal of Advances in Medical Education & Professionalism

The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations

KENNETH D. ROYAL[1]*, MYRAH R. STOCKDALE[2]

[1]*Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA;* [2]*College of Pharmacy and Health Sciences, Campbell University, Buies Creek, NC, USA*

Abstract

Introduction: Research has asserted MCQ items using three response options (one correct answer with two distractors) is comparable to, and possibly preferable over, traditional MCQ item formats consisting of four response options (e.g., one correct answer with three distractors), or five response options (e.g., one correct answer with four distractors). Some medical educators have also adopted the practice of using 3-option responses on MCQ exams as a response to the difficulty experienced in generating additional plausible distractors. To date, however, little work has explored how 3-option responses might impact validity threats stemming from random guessing strategies, and what impact 3-option responses might have on cut-score determinations, particularly in the context of medical education classroom assessments. The purpose of this work is to further explore these critically important considerations that largely have gone ignored in the medical education literature to this point.

Methods: A cumulative binomial distribution formula was used to calculate the probability that an examinee will answer at random a given number of items correctly on any exam (of any length). By way of a demonstration, a variety of scenarios were presented to illustrate how examination length and the number of response options impact examinees' chances of passing a given examination, and how subsequent cut-score decisions may be impacted by these factors.

Results: As a general rule, classroom assessments containing fewer items should utilize traditional 4-option or 5-option responses, whereas assessments of greater length are afforded greater flexibility in potentially utilizing 3-option responses.

Conclusions: More research on items with 3-option responses is needed to better understand what value, if any, 3-option responses truly add to classroom assessments, and in what contexts potential benefits might be discernible.

Keywords: Medical education; Assessment; Evaluation; Psychometrics

**Corresponding author:*
Kenneth D. Royal,
Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA,
Tel: +1-919-5136100
Fax: +1-919-5136464
Email: kdroyal2@ncsu.edu

Please cite this paper as:
Royal KD, Stockdale MR. The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. J Adv Med Educ Prof. 2017;5(2):84-89.

Received: 30 August 2016
Accepted: 14 December 2016

Introduction

Research by Rodriguez (1) has asserted MCQ items using three response options (one correct answer with two distractors) is comparable to, and possibly preferable over, traditional MCQ item formats consisting of four response options (e.g., one correct answer with three distractors), or five response options (e.g.,

one correct answer with four distractors). Some medical educators have also adopted the practice of using 3-option responses on MCQ exams as a response to the difficulty experienced in generating additional plausible distractors. While it would seem solving the problem of creating a third or fourth plausible distractor could be fairly easily alleviated by tapping the expertise of one's colleagues, advanced students/residents, etc. (2-4), some remain committed to administering items with 3-option responses. While there is some research (and continually budding interest from faculty) supporting the use of items with 3-option responses, we must consider what effects this might have on guessing strategies and cut score determination decisions so as to avoid any unintended consequences (5). Thus, the purpose of this work is to further explore these critically important considerations that have largely gone ignored in the medical education literature to this point.

Overview of guessing

Guessing on multiple-choice examinations (MCQs) has long been recognized as a serious threat to score validity. Psychometricians generally have recognized three types of guessing: 1) informed guessing; 2) cued guessing; and 3) random guessing (6, 7). While most educators recognize informed and cued types of guessing behaviors are common on medical school examinations, most educators have largely dismissed the notion that one could attain a desirable score result while relying purely on random guessing. After all, on the surface this appears to be a very poor strategy for completing an examination. However, let us look more closely at the impact of this particularly unassuming type of guessing behavior.

With random guessing, an examinee's odds of correctly answering any given item increase with fewer response options. For example, for someone guessing purely at random, s/he would have a 1 in 5 (20%) chance of correctly answering an item with 5 response options; a 1 in 4 (25%) chance of correctly answering an item with 4 response options; and a 1 in 3 (33%) chance of answering an item with 3 response options. One may calculate the probability that an examinee will answer a given number of items correctly on any exam (of any length) by using the cumulative binomial distribution in Excel. The formula for calculating these probabilities is presented below:

P=(1-BINOM.DIST (*A*,*B*,*C*,TRUE))+BINOM. DIST (*A,B,C*,FALSE)

Where, *P* is the probability of an examinee getting the score of interest (or higher) by randomly guessing during a single administration, *A* is the score of interest (e.g., the number of correct responses), *B* is the length of the exam, and *C* is the probability of getting an item correct when randomly guessing (e.g., 1 divided by the number of response options).

An illustrative example

To illustrate why random guessing is a problem that medical educators should take more seriously, especially when cut scores are used, let us consider a hypothetical example. Imagine a medical education scenario in which students complete an exam consisting of 10 items. The instructor opted to administer only 10 items because the exam is considered particularly rigorous due to challenging items assessing very complex material. Further, the instructor opted to utilize a 3-option format for responses based on literature suggesting 3 options are largely comparable to 4 options when it is difficult to generate an effective third distractor. Now, suppose the instructor conducted a standard setting exercise with colleagues and determined the following raw score performance ranges were appropriate and defensible: 0-4=Fail; 5-7=Pass; 8-10=Honors. In such a scenario, we used the cumulative binomial distribution formula presented previously to determine the probability that a student utilizing a random guessing strategy would attain a given number correct (Table 1).

This example demonstrates that a student guessing completely at random would have a .2064 (about a 20.64%) probability of passing this exam. Most would agree that this probability for achieving a 'Pass' verdict is much too large, thus the current cut score between a Pass/Fail decision must be revised. With most education research it is customary to use a 95% confidence level, meaning education researchers are willing to accept a 5% chance of making an error. In this example, this would mean we must move the minimum cut score from 5 (50% correct in order to attain a Pass verdict) to 7 (70% correct in order to attain a Pass verdict) in order to ensure that no examinees would be likely to pass the exam by simply guessing at random. As stated previously, because this exam is particularly difficult a shift of the cut score from 5 to 7 would likely be indefensible. This creates quite the dilemma for a medical educator wishing to administer a rigorous assessment while maintaining fair and defensible performance standards.

Now, let us look at the same example using 4- and 5-option responses. In this scenario, let us assume the instructor was able to generate three and four plausible distractors, respectively,

Table 1: Probability of attaining a given raw score with 3 response options on a 10-item test

Number correct	Probability of attaining number correct	Performance category
1	0.9818	Fail
2	0.8920	
3	0.6930	
4	0.4316	
5	0.2064	Pass
6	0.0732	
7	0.0185	
8	0.0032	Honors
9	0.0003	
10	0.0000	

for each of the 10 items. The probability that a student utilizing a random guessing strategy would attain a given number correct is presented in Table 2.

In this scenario, a student guessing completely at random would have approximately a.07 (or about an 8%) probability of passing this exam when each item consisted of 4-response options, and about a .03 (about 3%) probability of passing when each item consisted of 5- response options. In both instances, the probability of passing the exam is much smaller than the exam consisting of 3-response options. Interestingly, though, the exam consisting of 4-response options still produces a probability of passing that is greater than the typical 5% threshold we would prefer. If using 4-response options, the instructor would likely still need to revise the cut score in this scenario. The result would be a move from 5 (50% correct in order to attain a Pass verdict) to 6 (60% correct in order to attain a Pass verdict) in order to ensure that examinees would be unlikely to pass the exam by simply guessing at random. Although this shift in the cut score seems less egregious than the shift given 3-response options, it still may be problematic and indefensible in such a scenario. So, what exactly does this mean for medical educators conducting routine classroom assessments?

Implications for medical educators

While there is research and a continual budding of support from medical educators to support the use of MCQ items with 3 response options, it is critical that one remains cognizant of what effect this will have on guessing strategies, and on cut score determination decisions. Simply put, exams consisting of items with fewer response options will increase the likelihood that examinees utilizing a random guessing approach will pass that exam. The choice of using 3, 4 or 5 option responses also has significant implications with regard to setting cut scores and various performance standards. Most medical school classroom assessments carry moderate-to-high stakes for students, thus it is imperative that the placement of a cut score is at a position in which it is highly unlikely an examinee could achieve by random guessing. Of course, any decision about what constitutes "highly unlikely" is arbitrary, but given the conventions of education research a level of 5% seems reasonable for classroom assessments. Table 3 provides recommendations for cut score thresholds for the minimum performance category (e.g., the threshold for a Pass/Fail decision) relative to exam length (from 10 to 100 items, by increments of 5) and the number of item response options.

A visual scan of Table 3 illustrates the impact random guessing may have on minimum

Table 2: Probability of attaining a given raw score with 4- and 5- option responses

Number correct	Probability of attaining number correct with 4-response options	Probability of attaining number correct with 5-response options	Performance category
1	0.9437	0.8926	Fail
2	0.7560	0.6242	
3	0.4744	0.3222	
4	0.2241	0.1209	
5	0.0781	0.0328	Pass
6	0.0197	0.0064	
7	0.0035	0.0009	
8	0.0004	0.0001	Honors
9	0.0000	0.0000	
10	0.0000	0.0000	

Table 3: Minimum recommended raw cut score placements for items with 3, 4, and 5 options

Exam length (# of items)	Minimum cut score with 3 response options	Minimum cut score with 4 response options	Minimum cut score with 5 response options
10	7	6	5
15	9	8	7
20	11	9	8
25	13	11	9
30	15	13	11
35	17	14	12
40	19	16	13
45	21	17	15
50	23	19	16
55	25	20	17
60	27	22	18
65	27	23	19
70	31	25	21
75	33	26	22
80	34	27	23
85	36	29	24
90	38	30	25
95	40	32	27
100	42	33	28

performance category threshold decisions. The table illustrates the number of items that would need to be answered correctly in order to be confident within a 5% error tolerance that the response pattern was not due to random guessing. Specifically, the fewer response options provided, the greater the minimum (raw) cut score need be. For example, an exam consisting of 50 items would require a minimum (raw) cut score of 23 if the items consisted of 3 response options, and a minimum (raw) cut score of 16 if the items consisted of 5 response options to ensure examinees could not achieve the lowest meaningful performance category by random guessing given a 5% maximum error tolerance. The difference of requiring correct answers to 7 additional items when fewer options are presented offers a number of additional considerations for medical educators. Specifically, what impact might the decision to use items with 3-response options relative to traditional item formats with 4- or 5-response options have on setting fair and defensible standards? Clearly, fewer response options equate to a higher minimum cut score, but what effects might this have on the substantive meaning of that minimum cut score? What additional unintended consequences might result from using fewer response options?

Other considerations

Practically speaking, MCQs with 3-options are typically easier to write because they require development of fewer plausible distractors than 4- and 5-option MCQs. Plausible distractors are those incorrect options that relate to the objective being assessed. Educators should recognize that implausible distractors add to test error and therefore are not contributing meaningfully (in fact, they are doing quite the opposite). A good distractor targets learners' insufficiency in the content or competency domain being assessed. Item distractors are often written to diagnose common misconceptions (e.g., order of operations) about a topic. It benefits medical educators when others, such as fellow content experts or residents, review items and their distractors for plausibility.

In addition to distractor plausibility concerns, the amount of time per item can be problematic when trying to assess multiple content or competency domains. Examinees sometimes run out of time and render random guesses in hopes of gaining additional points. When this occurs, scores are contaminated with error and the resulting measure of student performance will be inaccurate. Proponents for 3-option responses contend examinees will require less time to complete each item, and may lead to more valid score results.

Alternatively, modifying exams in which a significant number of students have historically run out of time could give better estimates of difficulty and reliability than previous versions. However, in reducing the number of distractors that an item has, there is usually a small decrease in each item's difficulty and the exam's overall reliability (1). This translates to a need for more high-quality items per exam to compensate for

diminished difficulty and reliability.

It is well-documented that offering a mix of easy, moderate, and difficult items tends to increase reliability estimates (8). However, medical education classroom assessments are notorious for suffering from low reliability estimates, largely due to over-representation of easy items ($p \geq 0.75$). The extent to which items are easy due to their complexity, an artifact of student learning, instructional familiarity effects (9), or some other factor certainly must be considered as well. As Royal and Guskey (10) note, in the context of classroom assessment, when teaching is effective and learning is successful score results should resemble a negatively skewed distribution with few, if any, students receiving low marks. Thus, the importance of reliability as an indicator of quality is often deceiving under these circumstances and becomes a less important indicator of quality.

Discrimination coefficients are often helpful for informing instructors of whether or not students performed as expected. To that end, instructors can have some, albeit subjective, estimate of the degree to which the exam functioned as intended when administered to the given sample. It should be noted that discrimination coefficients simply need to be positive (e.g., 0.01 or higher), as opposed to the guidelines proposed for norm-referenced examinations (e.g., greater than 0.2, 0.3, or 0.4) which have largely been misunderstood by most medical educators (11). That is, because classroom assessments often yield high scores for examinees and items difficulty estimates tend to indicate most items are "easy", the discrimination coefficient becomes less important in this context than it would be in a norm-referenced assessment scenario (e.g., MCAT exam, etc.) (10, 12).

Exam length is another important consideration, as having too few items creates range issues and inflates error estimates. Therefore, educators need to ensure there are enough items to adequately measure students' abilities, while simultaneously ensuring someone cannot pass an exam by utilizing a random guessing strategy. Research by Fisher (13) indicated exams should consist of a minimum of 25 items, when possible, because measures produced with this many items yield error estimates that generally are stable while reliability estimates can still reach levels exceeding 0.90. As a general rule, the more items administered the more statistically stable the scores will be; however, at some point the cognitive load for examinees may become too great or examinees may have insufficient time, so these factors need to be appropriately considered and properly accounted for when determining an appropriate number of items to administer (14).

Of course, some individuals familiar with psychometric models might argue the solution for modeling guessing is simple: use a 3-parameter logistic (3-PL) item response theory (IRT) measurement model to account for a pseudo-guessing parameter to adjust students' scores. While this approach is certainly a possibility for individuals working in large-scale, medical licensure and certification testing environments, it is entirely inappropriate for classroom assessments as 3PL models require a minimum sample size of 1,000 examinees (15-17). Therefore, corrections for guessing using 3-PL IRT models is not a viable option for medical education classroom assessments. The use of Rasch measurement models, however, is a possibility for classroom assessments (11) and may offer insights about guessing (18, 19).

Conclusion

While there is some research to support the use of items with 3-option responses, we must remain cognizant of what effect this will have on guessing strategies and cut score determinations. As demonstrated here, the probability of attaining an invalid score result tends to increase in scenarios involving a small number of items. Much more research on items with 3-option responses are needed to better understand what value, if any, 3-option responses truly adds to assessments, and in what contexts potential benefits might be discernible.

Conflict of Interest: None declared.

References

1. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas. 2005;24(2):3-13.
2. Denny P, Luxton-Reilly A, Hamer J. The Peer Wise system of student contributed assessment questions. In Proceedings of the tenth conference on Australasian computing education-Volume 78. Australia: Australian Computer Society, Inc; 2008. p.69-74.
3. Haladyna TM. Developing and validating multiple-choice test items. USA: Routledge; 2012.
4. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas. 1993;53(4): 999-1010.
5. Royal KD, Puffer JC. The consequential validity of ABFM examinations. J Am Board Fam Med. 2014;27(3):430.
6. Rogers HJ. Advances in Measurement in Educational Research and Assessment. Oxford, UK: Pergamon; 1999. p.235-43.
7. Royal KD, Hedgpeth MW. A novel method for evaluating examination item quality. Int J Psychol Stud. 2015;7(1):17-22.

8. Royal KD, Hecker K. Understanding reliability: A review for veterinary educators. J Vet Med Educ. 2016;43(1):1-4.
9. Royal KD, Hedgpeth MW, Smith KW, Kirk D. A method for investigating "instructional familiarity". Ann Med Health Sci Res. 2015;5(6):428-34.
10. Royal KD, Guskey TR. On the appropriateness of norm- and criterion-referenced assessments in medical education. Ear Nose Throat J. 2015;94(7):252-4.
11. Royal KD, Gilliland KO, Kernick ET. Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. Anat Sci Educ. 2014; 7(6):450-60.
12. Royal KD, Guskey TR. The perils of prescribed grade distributions: What every medical educator should know. J Contemp Med Educ. 2014;2(4):240-1.
13. Fisher WP. The cash value of reliability. Rasch Meas Trans. 2008;22(1):1160-3.
14. Royal KD, Hedgpeth MW. Balancing test length with sufficiently reliable scores. Educ Med J. 2015;7(1):64-6.
15. Hulin CL, Lissak RI, Drasgow F. Recovery of two- and three parameter logistic item characteristic curves: A Monte Carlo study. Appl Psychol Meas. 1982;6(3):249-60.
16. Lord FM. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educ Psychol Meas. 1968;28(4):989-1020.
17. Yen WM. A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika. 1987;52(2):275-91.
18. O'Neill TR, Royal KD, Puffer JC. Performance on the American Board of Family Medicine (ABFM) certification examination: are superior test-taking skills alone sufficient to pass? J Am Board Fam Med. 2011;24(2):175-80.
19. Royal KD, O'Neill TR. Using the CUTLO procedure to investigate guessing. Rasch Meas Trans. 2011;25(1):1319-20.