THE PREVALENCE OF ITEM CONSTRUCTION FLAWS IN MEDICAL SCHOOL EXAMINATIONS AND INNOVATIVE RECOMMENDATIONS FOR IMPROVEMENT

***Kenneth D. Royal,[1,2] Mari-Wells Hedgpeth[1]**

1. Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, North Carolina, USA
2. Department of Family Medicine, UNC School of Medicine, University of North Carolina at Chapel Hill, North Carolina, USA
**Correspondence to kdroyal2@ncsu.edu*

Disclosure: The authors have declared no conflicts of interest.
Received: 16.08.16 **Accepted:** 27.09.16
Citation: EMJ Innov. 2017;1[1]:61-66.

ABSTRACT

The purpose of this study was to calculate the prevalence and nature of item construction flaws within one large medical school and to identify several innovative approaches that may serve as potential remedies for these problems. Results indicated that approximately one in five items contained a construction flaw, with the overwhelming majority of flaws involving poor quality distractors. A series of innovative recommendations are presented, including modern psychometric analytical techniques to more thoroughly inspect data, item manipulation techniques, and the use of innovative item types that may alleviate the need for distractors altogether.

Keywords: Multiple-choice questions (MCQs), item writing, item quality, assessment, medical education, educational measurement, psychometrics, testing, innovative items.

INTRODUCTION

Multiple-choice questions (MCQs) continue to be the preferred method of assessment in medical education due to the ease of administration and scoring, especially with large class sizes. Teaching faculty members are typically tasked with the challenge of developing items for classroom assessments. However, because the stakes associated with these assessments typically are moderate-to-high in nature, the need for quality items, particularly in terms of their construction, is paramount because items with construction flaws introduce measurement errors that threaten the validity of students' performance measures. Fortunately, item construction flaws can largely be mitigated with careful attention, by following best practice guidelines, and by making use of innovative item types. To that end, we sought to: i) calculate the prevalence of construction flaws at one medical school, ii) characterise the nature of these flaws, and iii) identify some innovative approaches that would likely mitigate many, if not most, of these flaws.

BACKGROUND

MCQs are the most commonly utilised assessment method used in medical education classroom assessments. This largely is due to the ease of administration, more objective and transparent scoring processes, and increased defensibility of scores. Many credentialing organisations, such as the United States Medical Licensing Examination® (USMLE), and various subspecialty boards comprising the American Board of Medical Specialties (ABMS), largely depend on MCQs for assessing their future and current workforce. Given the prevalence of testing with MCQs in medical education, it is important that item authors be aware of the major principles of sound item construction.[1] Considering that assessment comprises a significant amount of educators' time,[2] the teaching faculty should be provided with the opportunity to learn the principles of item writing.

Jozefowicz et al.[3] reported that teaching faculty are not routinely trained on how to develop quality

MCQs. As a result, items that are authored by teaching faculty do not always meet the recommended item writing criteria that have been established and widely circulated by experts in the field.[1,4-7] More specifically, items that contain technical flaws may contaminate examinees' scores with errors that interfere with both the accuracy and the valid interpretation of exam results.[8-11] It is imperative to evaluate the degree to which item flaws exist in a medical school's pooled item bank because inferences made about score results typically carry moderate-to-high stake implications for students (they are used to determine class rank and promotion to the next programme year and to identify suitable candidates for a residency programme, for example).

At a large public medical school in the southeast of the USA that offers Doctor of Medicine degrees, considerable resources are devoted to the pursuit of quality exam items. A team of assessment and testing experts, with significant experience in academia and the professional medical certification industry, work to ensure that faculty-generated items are sound in terms of both their technical quality and their psychometric properties (desirable reliability estimates and adequate discrimination indices, for example). All items appearing on exams are reviewed by this team and items flagged with technical flaws are reported to the faculty for potential revision. Furthermore, the assessment team routinely conducts workshops to educate the faculty regarding item writing, psychometric indicators, modern validity conceptualisations, and a host of other assessment-related issues. Given that so many resources and efforts are devoted to improving classroom assessments, it would be expected that the item bank at this institution would be particularly robust.

METHOD

Instrumentation

A systematic review of all preclinical (Year 1 and 2) MCQs presented on midterm and final exams was performed using an unpublished instrument developed by the late Prof Linnea Hauge (Table 1) and adapted from Haladyna et al.[1] who provided guidelines (as opposed to 'hard and fast' rules) intended to maximise item clarity and minimise validity threats stemming from various sources of error (an examinee's 'testwiseness' skills, construct irrelevance variance, for example).

All items were reviewed by two assessment experts. The instrument essentially collapsed the most prevalent item construction errors into one of eight flaw types and provided assessors with the ability to easily tally the number of flawed items. The two assessors worked together to read, review, and classify items as containing an item construction flaw(s) or as meeting the item writing standards of Haladyna et al.[1] Flawed items were identified as any item in which the item writer ignored one or more of the principles of quality item development. While each item may have had more than one technical flaw, for the purpose of scoring items in this study, the experts coded only one flaw per item; this was the flaw that, in their agreed opinions, was the most severe violation of the set standards. In these instances, secondary flaws were noted in the comments section of the scoring rubric. Items not containing a construction error were considered to have met recommended item writing principles.

Rating Process

A total of 2,204 items were carefully read, discussed, and classified according to the wording of each item by two assessment experts. Instances in which the experts disagreed on the type of flaw were flagged and the item was reviewed again by both experts individually; each expert noted the reasoning based upon Haladyna et al.'s[1] recommendations. The experts then discussed their individual decisions and worked to find a consensus opinion. All disagreements by experts were due to items having more than one identified technical flaw. The number of items with construction flaws versus the number of items that met the guidelines were calculated.

Examination and Item Characteristics

All items appearing on mid-term and final exams from the first 2 years of the pre-clinical, undergraduate medical school curriculum during the 2012–2013 academic year were investigated. The disciplines assessed by these exams covered basic science courses (microbiology, anatomy, physiology, and immunology) in the first year (MS1) and the major organ systems (cardiovascular, respiratory, gastrointestinal, renal-urinary, endocrine, reproductive-genetics, brain, and musculoskeletal) in the second year (MS2). One to three weeks of instruction were covered on each of the exams analysed, depending on the length of the overall course, exam items were authored by multiple

faculty staff who taught in each course and were considered content experts for their respective subject area. Each course may have had a dozen or more lecturers, all of whom may have contributed exam items. Most exam items were of the single best answer format with either four or five answer options.

The number of items appearing on these exams ranged from 28–100. There were 182 students enrolled in the MS1 cohort, and 175 students enrolled in the MS2 cohort. The MS1 data set consisted of 17 exams, with each exam assigned a score depending on the percentage of items meeting the established and widely recognised guidelines for sound item construction.[1,4-7]

Table 1: Types of item construction flaws.

- Items written in the negative (e.g. "Which of the following are not true?" or "All the following are correct except:")
- Items that included the use of 'none of the above', 'all of the above', or that used combinations of answer options within the distractors (e.g. K-Type questions)
- Items that were unfocussed (e.g. did not ask a direct question or required the examinee to read all answer options before being able to answer)
- Items that had answer options that were not homogenous, not of equal length, or the correct answer repeated elements included within other options
- Items that included the use of extreme language (e.g. always or never)
- Items that were tricky (e.g. extraneous reading in the answer options that was not required, making the item unnecessarily complicated)
- Items that were not formatted properly (e.g. vocabulary was not appropriate, punctuation was not correct, or item formatted horizontally instead of vertically)
- Items whose distractors included the use of humour or were not plausible

Instrument adapted from unpublished data from Prof Linnea Hauge.

Table 2: First and second year exam descriptive statistics.

	MS1 mean (SD)	MS2 mean (SD)
Number of items used	56.72 (13.02)	57.20 (20.40)
Lowest exam score	59.32 (4.03)	58.57 (8.17)
Highest exam score	98.85 (1.15)	99.72 (0.57)
Exam score	84.08 (2.76)	85.09 (2.51)
Exam SD	7.69 (0.59)	7.97 (1.30)
Exam reliability (KR-20)	0.66 (0.08)	0.67 (0.10)
Standard error of measurement	2.45 (0.38)	2.41 (0.48)

MS1: first year; MS2: second year; SD: standard deviation; KR-20: Kuder–Richardson Formula 20.

Table 3: Frequency and type of flawed items by programme year.

Technical flaws	MS1 exams (n=1,034)	MS2 exams (n=1,170)	Total flaws (n)
Negatives used in stem or distracters (e.g. except, not true, least likely)	27	61	88
None of the above/all of the above, K-type, true-false	93	36	129
Unfocussed stem	97	25	122
Length of distracters is unequal, not in logical order	60	50	110
Grammatical structure and/or extreme language	4	2	6
'Tricky'	0	1	1
Inappropriate vocabulary and/or language	2	2	4
Distractors that are not plausible, use of humour	2	1	3

MS1: first year; MS2: second year.

The MS2 data set consisted of 20 exams, with each exam assigned a score regarding the percentage of items meeting these same recommended guidelines. Table 2 presents descriptive statistics for MS1 and MS2 exams. All exams were administered via a standardised web-based assessment system with a secure browser to mitigate sources of error stemming from conditions of administration.[9] Students were allotted approximately 1 minute and 40 seconds per question on average.

RESULTS

The percentage of exam items meeting guidelines for the MS1 courses was found to be 72.43%, and 84.79% for MS2 courses. Of the 2,204 total items administered to students during the 2012–13 academic year, 463 (21.01%) items contained flaws (Table 3). The most frequent item writing flaws found across both MS1 and MS2 courses was 'none of the above/all of the above', combinations of answer options (K-type), or true/false formats (n=129), followed by unfocussed stems (n=122), uneven formats of answer options where the correct answer was the longest option (n=110), and the use of negatives in item stems or distracters (n=88). Other types of flaws, such as grammatical structure, inappropriate language, implausible distracters and the use of humour, and tricky items were far less frequent, with 14 collective occurrences.

DISCUSSION

Substantive Results

The exam items reviewed were representative of all faculty-authored items administered during the first two preclinical programme years at the medical school. Results indicated approximately 79%, or about 1 in 5, of all items administered met the recommended guidelines for construction quality, while 21% did not. Given all the expert personnel, resources, and meticulous reviewing efforts provided to the faculty, we believe these values serve as a reasonable and potentially best case estimate for item construction flaws appearing on medical school classroom examinations.

On the surface, this finding is quite alarming as it suggests that about one in five items contain a source of error that could otherwise be mitigated with more careful discernment on the part of the faculty item writers.[8-9,11] It is important to note however, that there was considerable variation across course year. MS1 courses focussing on the basic sciences contained considerably more flaws (27.56%), with approximately 1 in every 3.62 items containing a technical flaw, whereas MS2 courses focussing on the clinical sciences contained considerably fewer flaws (15.21%), with approximately 1 in 6.57 items containing a technical flaw.

It is important to note that while the institution devotes considerable resources and training to help faculty generate items that are technically sound in construction, it is unknown exactly how many faculty staff take part in training exercises and/or use the resources made available to them. While it would be ideal to train every faculty member who contributes to the medical education enterprise, this simply is not realistic given the enormous number of medical school faculty members, often in the hundreds, and the many competing demands of the faculty, for whom education is often a lower priority.

Furthermore, it remains unknown how many faculty members take part in instruction and/or contribute items to exams. Given course directors often have different styles for managing courses therefore any answers given are likely to be highly variable. Of course, it is hoped that responsible course directors will ensure continual efforts are made each year to improve items and over time, this should result in a significantly improved item bank. However, we are fearful that such continuous improvements may not be entirely realistic. For example, a best practice in testing recommends faculty staff alter their exams each year as a preventative measure to combat cheating, as students often share information about items appearing on exams.[12] When items are replaced with new ones, it is unlikely that the new items are any better in terms of construction quality, especially if the items were generated as last minute substitutes which faculty staff acknowledge is often the case. Of course, the extent to which faculty staff heed recommendations about improving their exams also remains unknown. We suspect this practice is also highly variable and likely depends on many factors, not the least of which is the depth of one's item bank and one's true commitment to conducting objective assessments.

With respect to the consequences that may result for students, this also remains largely unknown. On the one hand, it may be argued that students

significantly benefit from construction errors such as choosing the longest response option as this 'testwiseness' strategy is widely taught to students as a cued-guessing strategy when the answer is unknown. In such instances, students' performance measures will be inflated and an overestimate of what students truly know (or can do) will be obtained. On the other hand, some item construction flaws may work to the detriment of students. For example, a question that asks students to identify the response option that is 'not true' or 'least likely' may cause some students who truly understand the concept(s) in question to render an incorrect response. In such cases, students' performance measures will be deflated and underestimate what students truly know (or can do). In any instance, the mismeasurement stemming from these sources of error no doubt results in some students appearing more/less knowledgeable (or capable) than they actually are. From an assessment perspective, this is most unfortunate because the errors stemming from item construction are largely preventable by following the well-recognised guidelines for quality item construction and properly acting upon the findings generated from a review of psychometric (statistical) indicators.

Recommendations for Improvement

Several clever and easy-to-implement techniques exist to help item writers improve traditional MCQ items. For example, team item writing by way of leveraging the expertise of peers, residents, and interns can help generate additional plausible distractors. Another technique is 'nudging' and 'shoving'[13] where distractors are easily manipulated to alter an item's difficulty level. Some research also suggests that moving from the traditional four or five option responses to three options might alleviate the challenges of generating more than two plausible distractors without affecting student performance measures.[14] Options also exist with respect to scoring. For example, Rasch measurement models have proven to be very robust for medical education examinations.[15] These models investigate an examinee's response pattern relative to an expected structure based on a given set of items with varying degrees of difficulty. These analyses can provide useful insights regarding aberrant responses, problematic items, potential for guessing, etc.

If item writers administer electronic exams, then several innovative options noted recently in the psychometrics literature are possible (audio items provide one possibility, for example). Although research on the use of audio exams is currently sparse, the concept seems promising in some situations. Psychology research indicates that sounds are processed differently by the brain than visual information,[16] so it is possible that audio items may unlock improved measurements of students' knowledge, skills, and abilities. Advantageously, audio items are essentially a higher level of simulation compared with written MCQs (low fidelity simulation). For example, imagine a cardiovascular and/or respiratory item that presents the examinee with an audio file of the pertinent findings (e.g. heart arrhythmia, murmur, abnormal breathing associated with bronchitis, etc.) and asks the examinee to diagnose it. One challenge to this approach would be that exam administrators must stringently vet headphone/laptop activity for exam security purposes.[12]

'Hotspot' items provide another powerful option. These item types provide a graphic and allow examinees one click on the image to indicate the correct answer.[17] This item type alleviates the need to generate written distractors, as a click on any area outside the designated correct zone (on the graph) is incorrect. An example might include asking examinees to identify with one mouse click a particular vessel on an anatomy exam. 'Drag and drop' items are particularly helpful for mid-level simulation activities. For example, in a typical anatomical practical exam the student is asked to identify body parts by placing a flag on a specific location. The drag and drop electronic format could closely resemble this procedure and remove many of the challenges associated with practical exams (scheduling, time commitment, and cadavers, for example).

'Figural structured response'[17] items essentially ask students to move around pieces on a graphic to demonstrate their knowledge. An example might include asking students to click on nerves that are responsible for movement of the bicep or testing reflexes. 'Alternate choice' items display several images and ask examinees to identify the most appropriate/best option.[18] For example, an examinee must evaluate four different cell/tissue/organ stains and determine which one would most likely be the microscopic finding that corresponds to the symptoms of a given disease. Again, this item does not require generating names of other diseases to use as potential distractors and it focusses the examinee on the problem to be solved

without generating hypothetical distractors that might be implausible if presented in written form. This format more closely resembles actual practice.

CONCLUSION

Findings resulting from a systematic review of medical school exam items revealed that approximately one in five items contain an item construction flaw and the overwhelming majority involve ineffective distractors or unfocussed stems. The aforementioned innovative item types present a number of potential remedies, as they would largely mitigate the use of distractors, and help item authors to focus questions on clinical reasoning skills (as opposed to recall of knowledge) while potentially providing a more accurate measure of knowledge, skills, and abilities, minimise 'testwiseness' strategies (detecting cues in how the item or its distractors are presented and sequencing cues where the response to one item can trigger a response to a previously administered item, for example), as well as better simulating medical practice. At present, innovative item types have not yet been thoroughly explored in medical education, thus future research should explore the benefits and challenges associated with these promising item types.

REFERENCES

1. Haladyna T et al. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Appl Meas Educ. 2002;15(3):309-34.

2. Stiggins R, Conklin N, "Initial Observations of Classrooms," Teachers' Hands: Investigating the Practice of Classroom Assessment (1992), Albany: SUNY Press, pp.32-53.

3. Jozefowicz RF et al. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156-61.

4. Downing SM, "Twelve Steps for Effective Test Development." Downing SM et al. (eds.), Handbook for Test Development (2006), Mahwah: Lawrence Erlbaum Associates, pp.3-25.

5. Downing SM, Haladyna TM. Test item development: validity evidence from quality assurance procedures. Appl Meas Educ. 1997;10(1):61-82.

6. Case S, Swanson D, "Technical Item Flaws," Constructing Written Test Questions for the Basic and Clinical Sciences (2002) 3rd edition, Philadelphia: National Board of Medical Examiners, pp. 19-29.

7. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 1989;2(1):37-50.

8. Downing SM. 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. Adv in Health Sci Educ. 2005;10(2):133-43.

9. Royal KD, Hecker KG. Understanding reliability: a review for veterinary educators. J Vet Med Educ. 2016;43(1):1-4.

10. Royal KD, Hedgpeth M. Balancing test length with sufficiently reliable scores. Educ Med J. 2015;7(1):64-6.

11. Royal KD, Hedgpeth M. A novel method for evaluating examination item quality. Int J Psychol Stud. 2015;7(1):17-22.

12. Royal KD et al. The "10 most wanted" test cheaters in medical education. 2016. Available at: https://cvm.ncsu.edu/wp-content/uploads/2016/04/10-Most-Wanted-Cheaters-DoCS-presentation.pdf. Last accessed: 6 October 2016.

13. Royal KD. Using the Nudge and Shove Methods to Adjust Item Difficulty Values. J Vet Med Educ. 2015;42:239.

14. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas. 2005;24:3-13.

15. Royal KD et al. Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. Anat Sci Educ. 2014;7(6):450-60.

16. Ballas J, "Delivery of information through sound," Kramer G (eds.), Auditory Display (1994), Reading: Addison-Wesley, pp.79-94.

17. Parshall CG, Harmes JC. Improving the Quality of Innovative Item Types: Four Tasks for Design and Development. Journal of Applied Testing Technology. 2009;10(1).

18. Haladyna TM, "Multiple-Choice Formats", Developing and Validating Multiple-Choice Test Items (1994), Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers, pp.35-57.