THE ASIA- PACIFIC SCHOLAR
MEDICAL AND HEALTH PROFESSIONS EDUCATION

ORIGINAL ARTICLE

Published online: 2 May, TAPS 2018, 3(2), 19-24
DOI: https://doi.org/10.29060/TAPS.2018-3-2/OA1049

The process of developing a rubric to assess the cognitive complexity of student-generated multiple choice questions in medical education

Rebecca Grainger[1], Emma Osborne[2], Wei Dai[1] & Diane Kenwright[1]

[1]Department of Pathology and Molecular Medicine, University of Otago Wellington, New Zealand; [2]Higher Education Development Centre, University of Otago, New Zealand

Abstract
Cognitively complex assessments encourage students to prepare using deep learning strategies rather than surface learning, recall-based ones. In order to prepare such assessment tasks, it is necessary to have some way of measuring cognitive complexity. In the context of a student-generated MCQ writing task, we developed a rubric for assessing the cognitive complexity of MCQs based on Bloom's taxonomy. We simplified the six-level taxonomy into a three-level rubric. Three rounds of moderation and rubric development were conducted, in which 10, 15 and 100 randomly selected student-generated MCQs were independently rated by three academic staff. After each round of marking, inter-rater reliability was calculated, qualitative analysis of areas of agreement and disagreement was conducted, and the markers discussed the cognitive processes required to answer the MCQs. Inter-rater reliability, defined by the intra-class correlation coefficient, increased from 0.63 to 0.94, indicating the markers rated the MCQs consistently. The three-level rubric was found to be effective for evaluating the cognitive complexity of MCQs generated by medical students.

Keywords: *Student-generated Multiple-choice Questions, Cognitive Complexity, Bloom's Taxonomy, Marking Criteria, Moderation of Assessment*

Practice Highlights

- Allow enough time for several cycles of moderation between markers, especially when the subject matter is complex. While other researchers have reported reaching a high level of inter-rater reliability swiftly, our research highlights that it can take time for teams to agree on a marking approach for complex, clinically-based questions.
- Guide students to write questions that require the information in the full stem to answer the question. We found that without additional guidance, students often wrote detailed clinical vignettes that were followed by straightforward recall-type questions.
- Minimise levels of complexity included in the rubric. We found three levels of complexity sufficient to make practical distinctions in the quality of students' questions.

I. INTRODUCTION

Multiple choice questions (MCQs) are used widely in assessing medical education. Well-constructed MCQs can be valid and reliable assessment tools. (McCoubrie, 2004; Schuwirth & Van Der Vleuten, 2004). From a practical perspective, they are also reusable, easy to administer and easy to grade. While a recognized drawback of MCQs is that they tend to test memorization rather than analytical thinking (Schuwirth & Van Der Vleuten, 2004; Veloski, Rabinowitz, Robeson, & Young, 1999), it is possible to construct MCQs that do test students' ability to apply knowledge and analyse problems (Khan & Aljarallah, 2011; McQueen, Shields, Finnegan, Higham, & Simmen, 2014; Palmer & Devitt, 2007). Given that students modify their study strategies in accordance with the complexity of thinking they anticipate needing to use in summative assessment (Biggs, 1999; Scouller & Prosser, 1994), one challenge for medical educators is to develop cognitively complex MCQs that will foster the kind of analytical reasoning that students will need in their medical careers.

One facet of improving MCQs is developing clear guidelines for items that require cognitively complex thinking as well as memorization. This requires a framework for classifying the thinking needed to answer MCQs. Bloom's *Taxonomy of Educational Objectives* (Bloom, 1956) and the subsequent revision of Bloom's taxonomy (Krathwohl, 2002) are popular starting points for classifying MCQ (Bates, Galloway, Riise, & Homer, 2014; Buckwalter, Schumacher, Albright, & Cooper, 1981; Khan & Aljarallah, 2011; McQueen et al., 2014; Palmer & Devitt, 2007; Rush, Rankin, & White, 2016). However, the majority of these papers tend to describe the process of rating MCQs using such a taxonomy very briefly, perhaps implying that the act of categorizing questions can be assumed to be intuitive and straightforward. Yet when we attempted to score MCQs using a Bloom-derived taxonomy, we initially found it difficult to translate a theoretical approach to cognitive complexity into a practical marking guide.

Medical students at the University of Otago were tasked with writing case-based MCQs for topics in pathology. The purpose of this task was to engage students in deep, clinically relevant learning in a way that also fulfilled their need for material that prepared them for the end-of-year MCQ examination (Grainger, Dai, Osborne, & Kenwright, 2017). Our research team then developed a rubric to evaluate the cognitive complexity of these student-generated MCQs, and this paper reports this process of rubric development. We initially found a high level of disagreement between markers as to how questions should be scored, evidenced by a low level of inter-rater reliability. Through analyzing the cognitive processes required to answer the questions and revising our marking criteria, we subsequently achieved a high level of inter-rater reliability. This paper argues that assessing MCQs for cognitive complexity based on existing taxonomies is an achievable task for a non-specialist team and reports our process of developing marking criteria as a model for other teams attempting a similar task.

II. METHODS

The student-generated MCQ approach was used in four modules (cardiovascular, central nervous system, respiratory and gastrointestinal) of an anatomic pathology course at the University of Otago. One hundred and six fourth-year medical students were enrolled in the PeerWise platform, in which students create MCQs and answer questions that their peers have created. (University of Auckland, 2016). For each topic, each student was required to create at least two MCQs similar to those found in their end-of-year exam, each comprising a stem (case scenario with question), one correct answer and three or four plausible distractors.

A rubric based on Bloom's Taxonomy for evaluating the quality of these MCQs was developed over three iterations. The highest level of Bloom's taxonomy, *synthesis*, was not included in the rubric as it is not applicable to a pre-defined task such as writing MCQs. In the first round of moderation, 10 out of 201 MCQs were randomly selected and independently rated by three markers. Results were then shared between raters, and one of the raters (EO) identified patterns of agreement and disagreement using summative content analysis of keywords and phrases that indicated the steps the respondent needed to undertake to answer the question (Hsieh & Shannon, 2005). Following this analysis, MCQs that were representative of issues the markers disagreed were circulated between the team members. These questions were used as a starting point for structured conversations where each rater described the process that they had used to mark to the question. Then a subset of 15 out of 331 MCQs, followed by a further 100 out of 678 MCQs were rated, analysed and discussed in the same manner. After each round of moderation, the inter-rater reliability was determined by calculating the intra-class correlation coefficient (ICC) (Bartko, 1976). Three staff participated the rating process. One had content expertise (RG), while the other two had backgrounds in higher education (WD, EO). The project had ethical approval from the University of Otago Human Ethics Committee (D16/423).

III. RESULTS

After the first round of marking, there was a low level of inter-rater reliability (ICC = 0.543, 95% CI -0.668-0.912), suggesting raters were inconsistent in assigning the MCQs to levels in the six-level rubric. There was high level of agreement among raters about whether certain types of question should be classified as cognitively complex or not. For example, all raters marked questions requiring *recall* or *comprehension* of factual knowledge lower than questions required the respondent to make a diagnosis based on a clinical scenario (see Figure 1). However, raters were inconsistent on which level a question should fall within the low-order thinking category (i.e. *recall* or *comprehension*) and within the high-order thinking category (i.e. *application*, *analysis* or *evaluation*). As the aim of the task was to foster cognitively complex questions, we condensed Bloom's *recall* and *comprehension* levels into a single level. In line with literature indicating that *analysis* and *evaluation* frequently overlapped (Moseley et al., 2005) we condensed these two categories into one level, while retaining the distinction between *application* and *analysis*.

The inter-rater reliability slightly increased in the second round of marking using the simplified rubric (ICC = 0.62, 95% CI 0.105-0.869). Content analysis of characteristics of inconsistently marked MCQs showed that marking varied for clinical case-based MCQs. Some MCQs had recall-based questions nested within a stem that superficially featured a clinical case, and markers agreed after discussion that these should be treated as *recall/comprehension* questions (see Figure 2).

Which of the following is not a feature of infiltrating astrocytomas?

A. It accounts for around 80% of adult primary brain tumours.
B. High grade lesions have leaky vessels that exhibit contrast enhancement on imaging.
C. The transition from normal to neoplastic cells is indistinct.
D. Microscopically psammonoma bodies can be seen.

Marker comment: This question lacks a clinical scenario that would require the respondent to *apply* their knowledge to a real-life problem. To answer the question, the respondent needs to *recall* factual information associated with the condition and to *understand* aspects of the condition's appearance.

Figure 1. Question testing recall/comprehension without a clinical case in the stem
Note: Questions have been lightly edited for clarity and brevity (abbreviations expanded and extraneous description removed) but otherwise left as written by the students, reflecting understanding of pathology at a fourth year medical student level. Author's chosen correct answer is indicated in italics.

A 27-year-old man is rushed into the Emergency department after suddenly collapsing during a marathon run. Upon examination, the patient is found to have a heart rate of 110 bpm, a blood pressure of 70/50 mmHg, respiratory rate of 24 breaths per minute and temperature 36.7° C. A CT scan is ordered, and show a diagnosis of an aortic dissection. Which one of the following statements is false?

A. Because the patient is hypotensive, the aortic dissection is likely to be a group B aortic dissection according to the Stanford classification.
B. A normal 12-lead ECG (not including the tachycardic rate) in this patient would be consistent with the diagnosis.
C. The young age of the patient suggests Marfan's syndrome is a possible factor.
D. A finding of a difference in blood pressure greater than 20 mmHg between the right and left upper limbs contradicts the diagnosis.

Marker comment: Although the question includes a clinical scenario in the stem, it does not require the respondent to use this information because the diagnosis is stated in the stem. The possible answers include statements that test *recall* and basic *comprehension* of facts associated with the condition.

Figure 2. Recall/comprehension question nested in a clinical stem

There was an unclear boundary between *application* and *analysis*/*evaluation*. In the subsequent discussion, we agreed that questions where the respondent needed to choose a diagnosis from a straightforward list of symptoms should be classified as *application* (see Figure 3).

A 20-year old New Zealand European male presents with a three-day history of macroscopic haematuria, low grade fever and loin pain. He is otherwise well. He experienced a similar episode of haematuria with no other symptoms about a year prior, which resolved spontaneously. His uncle had his gallbladder removed but his family is otherwise well. He not taking any regular medicines. Observations: HR 64, BP: 140/90, RR: 18, Temp: 37.6.

What is the most likely diagnosis and management?

A. Pyelonephritis. Provide supportive care and discharge.

B. Cystic cancer. Requires radical cystectomy. Refer to surgeons immediately

C. *IgA nephropahty. Discharge to outpatient clinic for biopsy, conduct immunofluorescence. Start ACE inhibitor if appropriate*.

D. Post strep glomerulonephritis. Start methotrexate immediately.

Marker comment: This question requires the respondent to *apply* their knowledge of the condition to make a likely diagnosis from signs and symptoms, then to *recall* appropriate treatment. The respondent could also answer the question by *excluding* incorrect combinations of conditions and treatments, which would draw on a subset of *classifying/categorizing*.

Figure 3. Question testing application of knowledge

We classified as *analysis/evaluation* questions which required the respondent to combine and interpret multiple forms of information or to anticipate other findings associated with a condition. For example, some questions required the respondent to predict likely test results from presenting symptoms, or combine and weight the importance of different sets of observations (see figure 4). Based on this discussion, specific explanations of each level of the simplified rubric in the context of medical education were generated and incorporated into the rubric (Table 1).

A high inter-rater reliability was shown in the third iteration using the simplified and redefined rubric (ICC= 0.89, 95% CI 0.845-0.923), suggesting that raters were assessing MCQs in a consistent way. Raters also reported improved time efficiency using the new rubric compared to the first two iterations.

Mr. S is a 53-year-old male who presents to you, his general practitioner, with lethargy for the last 6 months that he feels is out of the ordinary. He says his wife thinks his face is puffier than usual, and he has also developed some acne which he has not had since he was a teenager. He has also been experiencing shortness of breath at rest, and has had a persistent cough of for the last 3 months. He is a now a non-smoker but has a 20 pack year history. He has a BMI of 24, and has never had diabetes. You order a CXR which shows a central hilar mass. You refer him to Wellington hospital to get a biopsy which is examined by the pathologist. What would the expected microscopic findings be?

A. Hyperchromatic, pleomorphic, mitotically active glandular cells with areas of necrosis.

B. Small blue cells with little cytoplasm, crush artefact, and containing neurosecretory granules

C. Sheets of hyperchromatic, pleomorphic, mitotically active cells with keratin whorls.

D. Glandular tissue with goblet cell atrophy and neoplastic change.

Marker comment: The question requires the respondent to *analyse* and *combine* several sources of information (signs and symptoms, history and x-ray results) to form a possible diagnosis, then to *anticipate* and *interpret* the likely microscopic findings for this diagnosis.

Figure 4. Question testing analysis/evaluation of knowledge

Level	Corresponds to Bloom's Taxonomy	Description
Level 1	*Recall & comprehension*	Knowing and understanding facts about a disease, classification, signs & symptoms, procedures, tests.
Level 2	*Application*	Applying information about a patient (signs & symptoms, demographics, behaviours) to solve a problem (diagnose, treat, test)
Level 3	*Analysis & evaluation*	Using several different pieces of information about a patient to understand the whole picture, combining information to infer which is most probable.

Table 1. Rubric with categorization levels and explanations for the cognitive domain

IV. DISCUSSION

Student-generated, cognitively complex MCQs help prepare medical students for examinations which include these question types. This paper addresses the extent to which classifying questions by cognitive level is reliable, valid and practical. It also indicates a need for future research into how best to guide students in developing sophisticated MCQs.

We found our final rubric to be a reliable measure of question complexity, as evidenced by the high level of inter-rater reliability. The difficulties we found in drawing distinctions between levels of complexity were largely consistent with the challenges and possible solutions identified previously. For example, a lack of clarity in the top levels of Bloom's taxonomy reflects other work suggesting that modelling the higher order skills hierarchically may not be appropriate. One major revision of the taxonomy reverses the order of the upper levels (Krathwohl, 2002) and other critics have suggested that the differences between higher order skills are not clear cut and that ranking these skills is somewhat arbitrary (Moseley et al., 2005). While some have attempted to argue that MCQs can draw on thinking skills at all levels (Bloom, 1956; Young & Shawl, 2013), these appear to either: relate to questions that would only require evaluative thinking if reasoned from first principles in the exam rather than memorized (Young & Shawl, 2013); or be MCQs asked in relation to an extended problem rather than containing all the necessary information within the stem (Bloom, 1956). In developing our rubric, we selected levels of cognition similar to other researchers (Rush et al., 2016; Vanderbilt, Feldman, & Wood, 2013), although we combined *comprehension* with *recall* rather than *application*, as some others have done (Khan & Aljarallah, 2011; Palmer & Devitt, 2007). This suited our purposes in assessing a subject with a very strong applied component, where there was a crucial and clear difference between understanding the salient features of a condition and being able to apply that knowledge to a clinical scenario. The performance and utility of the rubric will need to be determined in other MCQ sets.

The difficulty we experienced in deciding how complex questions were does not appear to have been reported elsewhere; it is possible that this process is more difficult with highly involved clinical questions or that other authors have chosen not to focus on this area. One paper that does utilize Bloom's taxonomy in rating student-generated physics MCQs found a high level of inter-rater reliability in marking questions (Bates et al., 2014). Despite this, the authors do note a similar issue to us in that they comment that it was easier to rate lower-order questions than to make distinctions between *application* and *analysis*. Here it is likely that the subject material could influence the ease of marking. Bates et al. (2014) rated students' physics MCQs, and it may be that it was easier to identify, for example, whether single- or multiple-step mathematical calculations were required in these kinds of problems than identify the thought processes associated with clinical scenarios in our research.

In terms of the practicality of our rubric, we found that the clearly redefined rubric was effective in simplifying the rating process and reducing rating time. For non-content experts, the new rubric has enabled them to judge the level of cognitive effort at the same level as a content expert.

A final and not fully resolved question is how best to guide students in writing complex, scenario-based MCQs. Our larger research project found that students tended not to utilise theoretical guidance on using a model such as Bloom's Taxonomy in developing their MCQs (Grainger et al., 2017). We therefore intend to develop a more concrete, example-based scaffold for item-writing and assess whether students produce a similar quality of questions using this modified guidance.

V. CONCLUSION

Developing a valid and readily useable rubric to assess student-generated MCQs was achievable. A further task is to apply this rubric to new sets of questions to further test its performance and utility.

Notes on Contributors

Dr. Rebecca Grainger is an academic rheumatologist in Department of Pathology and Molecular Medicine of University of Otago Wellington. She is passionate about patient-focused care and medical education. She is responsible for the overall coordination and implementation of the study and assisted preparation for publication.

Emma Osborne is a professional practice fellow in student learning at the University of Otago. Her research interests include e-learning and teaching & learning in medical education. She initiated the process of rubric re-development, conducted qualitative analysis and was responsible for manuscript preparation.

Wei Dai is a research assistant in University of Otago. She is currently a Ph.D. candidate in Educational Psychology. Her research interest lies in the area of student engagement in technology-enhanced learning. She was responsible for the quantitative analysis and manuscript preparation.

Associate Professor Diane Kenwright is the Head of Department of Pathology and Molecular Medicine of University of Otago Wellington. She is a registered pathologist and an enthusiastic medical educator. She approved this research and assisted the preparation for publication.

Ethical Approval

This research has been approved by the Human Ethics Committee of University of Otago (Level B), reference number D 16423.

Acknowledgements

The authors would like to thank all the students in year four of University of Otago Wellington MBChB programme who participated the research.

Declaration of Interest

Authors have no conflicts of interest, including no financial, consultant, institutional and other relationships that might lead to bias.

Funding

This research had no specific grant funding from any funding agency.

References

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762-765. doi: http://dx.doi.org/10.1037/0033-2909.83.5.762.

Bates, S. P., Galloway, R. K., Riise, J., & Homer, D. (2014). Assessing the quality of a student-generated question repository. *Physical Review Special Topics - Physics Education Research*, *10*(2), 020105.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57-75.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook 1, cognitive domain*. London, Longman.

Buckwalter, J. A., Schumacher, R., Albright, J. P., & Cooper, R. R. (1981). Use of an educational taxonomy for evaluation of cognitive performance. *Academic Medicine, 56*(2), 115-121.

Grainger, R., Dai, W., Osborne, E., & Kenwright, D. (2017). *Self-generated multiple choice questions engaged medical students in active learning but not in a peer-instruction process*. Paper presented at the 14th Asia Pacific Medical Education Conference (APMEC), Singapore.

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research, 15*(9), 1277-1288.
doi: https://doi.org/10.1177/1049732305276687.

Khan, M. U. , & Aljarallah, B. M. (2011). Evaluation of Modified Essay Questions (MEQ) and Multiple Choice Questions (MCQ) as a tool for Assessing the Cognitive Skills of Undergraduate Medical Students. *International Journal of Health Sciences, 5*(1), 39-43.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice, 41*(4), 212-218.
doi: https://doi.org/10.1207/s15430421tip4104_2.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*(8), 709-712.
doi: https://doi.org/10.1080/01421590400013495.

McQueen, H. A., Shields, C., Finnegan, D. J., Higham, J., & Simmen, M. W. (2014). Peerwise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention. *Biochemistry and Molecular Biology Education, 42*(5), 371-381.
doi: https://doi.org/10.1002/bmb.20806.

Moseley, D., Baumfiled, V., Elliot, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. P. (2005). *Frameworks for thinking: A handbook for teaching and learning*. Cambridge: Cambridge University Press.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education, 7*, 49-49.
doi: https://doi.org/10.1186/1472-6920-7-49.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education, 16*(1), 250. doi: https://doi.org/10.1186/s12909-016-0773-3.

Schuwirth, L. W. T., & Van Der Vleuten, C. P. M. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974-979. doi: https://doi.org/10.1111/j.1365-2929.2004.01916.x.

Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education, 19*(3), 267-279.
doi: https://doi.org/10.1080/03075079412331381870.

University of Auckland. (2016). PeerWise. Retrieved from https://peerwise.cs.auckland.ac.nz/index.php

Vanderbilt, A. A., Feldman, M., & Wood, I. K. (2013). Assessment in undergraduate medical education: a review of course exams. *Medical Education Online, 18*, 1-5.
doi: https://doi.org/10.3402/meo.v18i0.20438.

Veloski, J., Rabinowitz, H., Robeson, M., & Young, P. (1999). Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Academic Medicine, 74*(5), 539-546.

Young, A., & Shawl, S. J. (2013). Multiple Choice Testing for Introductory Astronomy: Design Theory Using Bloom's Taxonomy. *Astronomy Education Review, 12*(1), 1-27.
doi: https://doi.org/10.3847/AER2012027.

*Rebecca Grainger
Senior Lecturer
Department of Pathology and Molecular Medicine
University of Otago Wellington
PO Box 7343, 23a Mein St
Newtown
Wellington South 6242
New Zealand
rebecca.grainger@otago.ac.nz
+64 4 385 5541