Pierson, E., & Chuong, T. (2014). *What about the women?* Tech.Coursera: Coursera. Retrieved October 9, 2015, from https://tech.coursera.org/blog/2014/03/08/what-about-the-women/

Sancho Vinuesa, T., Oliver, M., & Gisbert, M. (2015). MOOCS in Catalonia: Fueling innovation in higher education. *Educación XX1, 18*(2), 125-146.

Siemens, G., Gasevic, D., & Dawson, S. (Eds.) (2015). *Preparing for the digital university: a review of the history and current state of distance, blended, and online learning*. MOOC Research Initiative: Bill and Melinda Gates foundation report.

Yuan, L., & Powell, S. (2013). *MOOCs and Open Education: Implications for Higher Education*. JISC-Cetis, University of Bolton, UK.

What Questions are MOOCs asking? An Evidence-Based Investigation

Eamon COSTELLO[1], Mark BROWN[2] & Jane HOLLAND[3]

[1] National Institute for Digital Learning, Dublin City University, {eamon.costello, mark.brown}@dcu.ie

[2] Royal College of Surgeons in Ireland, jholland@rcsi.ie

Abstract

Multiple Choice Questions (MCQs) are a core building block of many MOOCs. In this exploratory study we analyze a sample of MCQs from a number of MOOCs and evaluate their quality. We conducted this analysis using a framework informed by a body of empirical research, which describes several common flaws that may occur in the way MCQs are written or phrased. Studies have shown that the presence of these flaws are likely to compromise the reliability and validity of tests containing these MCQs, potentially leading to poorer pedagogical outcomes. Through our study we contribute to the broad debate of whether MOOCs are a force that can enable enhanced and improved pedagogies or whether they will be susceptible to replicating existing poor pedagogies or practises at scale.

Keywords

MOOCs, Multiple Choice Questions, Tests, Quality

1 Introduction

Multiple Choice Questions (MCQs) are ubiquitous in education. They are present in all disciplines, but perhaps more so in STEM and more quantitative areas. They have a long and established history in medical education where an extensive body of literature exists regarding their use (SCHUWIRTH & VAN DER VLEUTEN, 2004). In addition to their use in high stakes terminal examinations, tests (or quizzes) incorporating MCQs are also frequently used in formative assessmenta. They may be used in conjunction with Classroom Response Systems, or within innovative peer assessment systems such as Peerwise, which enables the development of student generated learning tools and peer assessment via MCQs (DENNY, HAMER, LUXTON-REILLY & PURCHASE, 2008).

Currently, MCQs are a key component of many (x)MOOCs. The results that learners receive from these MCQ tests may contribute to their summative assessment grade, and so ultimately towards the certificate or credentials that they receive by MOOC completion or participation, whether this be formal or informal in nature. Given the important role that MCQs may play in MOOCs, the question then arises as to their psychometric quality.

There is a large body of research literature specific to the quality and psychometric properties of MCQ examinations. Two key concepts described within this evidence-base are reliability and validity. When we ask how reliable some measurement tool is we are essentially asking whether if we take several measurements with that tool under similar conditions we would get similar results. The overall reliability of an MCQ assessment, and the performance of individual items within, can be evaluated by models such as Classical Test Theory. In addition to evaluating the reliability of the overall test score, this theory also enables the evaluation of individual questions, by means of item analysis (DE CHAMPLAIN, 2010). This typically involves calculating parameters that indicate whether particular questions are of poor quality, such as item difficulty, or item discrimination (DE CHAMPLAIN, 2010). Such problems may be the result of flaws in the construction or writing of the MCQs (Downing, 2005). For example, various factors can affect the reliability (or repeatability) of an MCQ. If the question posed by an MCQ is incomprehensible to students then they will effectively have to guess – meaning the answers are random. An MCQ that is reliable gives consistent results and we can then ask whether those results are valid – i.e. whether it is testing something meaningful. As a further example, a valid instrument may be expected to discriminate between students of high ability and those of low ability. Therefore, an MCQ which is trivially easy, or too guessable, and which could result in all students getting the same answer regardless of inherent ability, might be described as being *invalid*, or unfit for the purpose of discerning a student's true ability.

There has been some exploratory research into the Quality of MOOCs (LOWENTHAL & HODGES, 2015; MARGARYAN, BIANCO & LITTLEJOHN, 2014). However, this research has not examined MCQs (a key component of MOOCs) in any way. In this study we sought to determine whether a sample of MOOC MCQs exhibited any of the commonly described item writing flaws. Our study makes an important contribution by addressing this gap in the MOOC research literature, and by exploring questions regarding the quality of MCQs in MOOCs. This issue is critically important if MOOCs are to fulfill aspirations to deliver formal learning that can contribute towards recognized awards.

2 Methodology

2.1 Sampling strategy and data collection

Our aim was to evaluate a range of multiple choice questions, sampled from MOOCs that were in English, and primarily in domains where the principal evaluator had expertise. Expertise can be important in determining certain criteria of question quality, however several criteria do not require in-depth expertise and many require none. A survey of existing platforms, aggregators and published research (MARGARYAN, BIANCO & LITTLEJOHN, 2014) revealed approximately 300 eligible courses for our purposes and 12 of these were selected at random from a weighted distribution of the relative spread across the platforms of EdE/X, Coursera, Futurelearn, Iversity and Eliademy. This resulted in 8 courses in the area of Computer Science and one each from Humanities, Medicine and Health, Psychology and Mathematics. Most courses were delivered in collaboration with universities partners, with the exception of two, one of which was from an individual and the second from a non-profit Institute.

Data collection was labor intensive and somewhat complicated, in that each question and all of its options had to be manually copied and pasted from the relevant MOOC quiz into a spreadsheet, which then acted as a data store. Moreover, the correct answer then needed to be determined – which in some cases proved a difficult task. It was originally intended to take ten questions randomly from each of the selected courses; however in some cases the correct answer to a given question could not be determined, or the information could not easily be extracted, and so ultimately it was not possible to collect ten questions from all courses selected. Therefore, in order to maintain our sample size, extra questions were collected from other courses so that in total 116 MCQs were collected (average 9.6 MCQs per MOOC), for which the correct answer (or answers) were determined and recorded (and by corollary the incorrect options). In all this resulted in the collection of 475 data points for analysis.

2.2 Procedure and Instrument

There are various frameworks and guidelines which may be used in order to evaluate the quality of MCQ items and examinations. Guidelines may range from simple five item rubrics (DENNY, LUXTON-REILLY & SIMON, 2009; PURCHASE, HAMER, DENNY & LUXTON-REILLY, 2010) to extensive manuals such as that from the US National Board of Medical Examiners (CASE & SWANSON, 2003). For our study we selected a tool which describes 19 item-writing flaws, and which has previously been used within the context of Health Professions Education (TARRANT, KNIERIM, HAYES & WARE, 2006). In addition to the benefit of utilising an existing, validated framework, we wished to use a tool with the potential to facilitate some comparability of our findings, albeit within a different context.

2.3 Data analysis

The full list of item-writing flaws is given in the results section below (Table 1), but for the purpose of describing our method of analysis, we can divide them into two broad categories: the first are those that can be calculated quite simply and the second are those that require the qualitative review of a human evaluator. Those that may be identified by means of simple calculation include; long correct option, option position and inclusion of options such as *"all of the above"*. For instance it has been shown that the longest option provided within an MCQ is frequently the correct one. This is due to a cognitive bias of (untrained) question writers who first compose the correct answer, which they may take due care in doing, and then later the distracters (incorrect options) which they spend less time and attention on. The length of the options was simply calculated by counting the character length of each programmatically. The number of the options and the number of correct options were computed in a similar manner, as was the position of the correct option. Once again, the available evidence-base suggests that the third (of four options) is most frequently the correct one. The strings of *"all of the above"* and *"none of the above"* were programmatically detected, as these options are considered flawed in the TARRANT, KNIERIM, HAYES & WARE (2006) framework.

The remaining thirteen items in our framework required the qualitative input of an evaluator to answer questions such as: "are the distracters plausible?", "is the question error free?", "is the language of the question ambiguous?" and so forth (Table 1). All selected MCQs and associated options (including distracters) were reviewed individually and evaluated against our framework to determine whether a flaw was present or not. Potential flaws not covered by our existing framework, and some other noteworthy features, were also recorded when observed, although they do not contribute to the results presented at this time.

3 Results

In total, 116 MCQs were reviewed within this study, and a total of 83 item writing flaws (errors) were detected. At least one error was present in 55 (47.4%) of the MCQs analysed, and 21 MCQs (18.1%) contained more than one error. When grouped by source, one MOOC was found to have only a single error within the ten MCQs sampled from it, but all the other courses selected for inclusion within our study demonstrated more than one error in their sampled MCQs. The most frequently occurring flaw observed in our dataset was the presence of Convergence Clues, which was detected 17 times, in 14.7% of MCQs (Table 1). This flaw may be seen in a few different forms, but in essence occurs when the correct answer includes the most elements in common with the other options, or distracters. This is due to novice question writers including facets or aspects of the correct component more frequently in the alternative options, when attempting to compose plausible distracters. Thereafter, "test-wise" students reading this question can then correctly guess the correct option as being that in which repeated components most frequently occur. Questions with more than one correct answer were identified in 10 instances (8.6% of MCQs), making this the second most common type of flaw observed. Nine occurrences were found of complex or k-type MCQs; these questions ask students to select from a range of possible combinations of correct responses, which can often be guessed by processes of elimination.

Table 1: Presence of MCQ Item Writing Flaws in 116 MCQs from 12 MOOCs

Item Writing Flaw	**Number detected**	**Percentage of Total**
Convergence clues	17	14.7%
More than one correct answer	10	8.6%
Complex or K-type question	9	7.8%
Question contains implausible distracters	8	6.9%
Ambiguous or unclear language in the question	5	4.3%
Question is asked in the negative	5	4.3%
Fill-in-the-blank question	3	2.6%
Problem is in the options and not in the question stem	3	2.6%
Word repeats in stem and correct answer	2	1.7%
"All of the above"	2	1.7%
"None of the Above"	1	0.9%
Unfocused question stem	1	0.9%
Logical cues in stem and correct option	1	0.9%
Vague terms used (sometimes, frequently)	0	0.0%
Absolute terms used (never, always, none, all)	0	0.0%
Gratuitous information in question stem	0	0.0%
Grammatical clues in sentence completion	0	0.0%
True/false question	0	0.0%

By contrast, some item writing flaws were not found at all in our selected MOOCs, such as the use of relative or absolute terms including the adverbs *"sometimes"*, *"frequently"*, *"always"* or *"never"*. Such terms can have different meanings to different people – even supposedly absolute terms such as *"always"* or *"never"* may be interpreted differently (Holsgrove & Elzubeir, 1998). Moreover, absolute terms are not recommended because question writers may not always be able to account for all circumstances (*"never"* might hold true today but not tomorrow). Additional flaws not found within our dataset were; gratuitous information in the question stem, grammatical clues in the question as to the answer, or true/false questions.

In addition to the above flaws, we analysed 103 of our sampled to see how frequently the longest option was also the correct option. The longest option was found to be correct more often than would be expected by chance, and this difference was significant (χ^2 [1, N = 103] = 12.28705, p = 0.000456).

In order to examine the position or distribution of correct options, we limited our analyses to those MCQs which had four options, which gave us a total of 73 MCQs. The distribution of correct options is demonstrated in Figure 1 below. We observed that the third option, or option C, was most frequently the correct one, occurring in 23 of our MCQs, or 32% of the time; however this was not statistically significant (χ^2 [1, N = 73] = 4.315068, p = 0.229391).

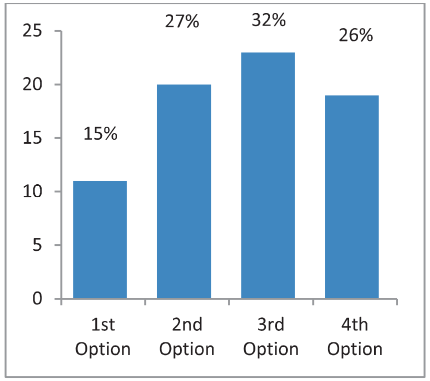

Figure 1: Frequency of Position of Correct Option in 73 Four Option MCQs

4 Discussion and Conclusion

This study sought to answer the question of whether MOOC MCQs exhibit commonly known item writing flaws, and we have demonstrated that this is indeed the case in nearly half of the questions sampled. Flawed MCQs may be confusing to examination candidates, particularly non-native speakers, and may reduce the validity of the examination process, penalizing some examinees (DOWNING, 2005; TARRANT, WARE & MOHAMMED, 2009).

Some item writing flaws were not detected within our dataset, or occurred infrequently; for example, although the third option was the one most often correct, as in previous studies, we did not find this to be statistically significant. However, absence of presence does not mean presence of absence, and the generalisability of this study could be improved upon by increasing the sample size, and analyzing a larger dataset. Likewise, formal evaluation and quality review of MOOC MCQs by established methods such as Classical Test Theory, might uncover additional flaws that are not immediately apparent to human evaluators without access to formal psychometric data or item analyses. These are two potential directions in which this research could be expanded.

We have demonstrated that these item writing flaws exist, and the question then arises as to their potential impact. For example, flawed items may fail to properly discriminate between students of high ability and those of low ability. Another potential impact is that students may fail a question simply because of the inherent fault in the question, such as a second correct option, rather than any error or lack of knowledge on their part. Alternatively, a student may “game” a test by guessing answers to questions with detectable flaws, simply by being “test-wise”, or aware of common grammatical errors or convergence clues. One study examining the impact of item writing flaws demonstrated that 33–46% of MCQs were flawed in a series of basic science examinations; the authors concluded that perhaps as many as 10–15% of the examinees were incorrectly graded as failing, when they should in fact have passed, due to the presence of these flawed items (Downing, 2002; Downing, 2005). Another study examined the quality of MCQs used in high stakes nursing assessments, and estimated that 47.3% of the MCQs reviewed were flawed (TARRANT & WARE, 2008). While the interaction between flawed items and student achievement can be complex, they demonstrated that borderline students benefited from these flawed items, which allowed a number of borderline students to pass examinations that they would otherwise have failed, had the flawed items been removed (Tarrant & Ware, 2008). In contrast, they also concluded that flawed items negatively impacted the high-achieving students in examinations, lowering their scores. If the MCQs from MOOCs analysed within our dataset were used in formal assessments, contributing towards credit or other attainment, it is plausible that a similar effect might occur on student achievement, with some students passing tests beyond their ability because of the presence of flawed items within the tests. However, within reliable and valid assessments scores should be an accurate reflection of the knowledge or skills they purport to examine.

A primary lesson that may be drawn from this study is the clear importance of proper training in MCQ writing. All question writers are prone to cognitive biases and errors, which proper training should alleviate but may not always overcome, and for this reason additional peer review and statistical analysis of MCQs is considered best practice. This is of course a time-consuming and expensive activity. Some very simple strategies could be included in MOOC MCQ engines to obviate obvious flaws (even as simple as ensuring question options are randomized which surprisingly few MOOCs seem to enforce). Many other common flaws could be detected through algorithmic means.

Within this study we simply counted the number of characters within each option in order to identify the longest one, but others have used computational techniques to look for the most linguistically complex option instead (BRUNNQUELL et al., 2011). Overall, it is hoped that this study will help remind stakeholders about the importance of a strong underlying pedagogy, supported by reliable and valid assessments. We believe that rich bodies of research exist that can help define, develop and ensure quality in our courses.

References

Brunnquell, A., Degirmenci, U., Kreil, S., Kornhuber, J., & Weih, M. (2011). Web-based application to eliminate five contraindicated multiple-choice question practices. *Evaluation & the Health Professions, 34*(2), 226-238. doi:10.1177/0163278710370459

Case, S. M., & Swanson, D. B. (2003). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia. PA: National Board of Medical Examiners Philadelphia.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008). PeerWise: Students sharing their multiple choice questions. Paper presented at the *Proceedings of the Fourth International Workshop on Computing Education Research,* 51-58.

Denny, P., Luxton-Reilly, A., & Simon, B. (2009). Quality of student contributed questions using PeerWise. Paper presented at the *Proceedings of the Eleventh Australasian Conference on Computing Education-Volume 95,* 55-63.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do Multiple-choice Item-writing principles make any difference? *Academic Medicine, 77*(10), S103-S104.

Downing, S. M. (2005). 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. *Advances in Health Sciences Education, 10*(2), 133-143.

Holsgrove, G., & Elzubeir, M. (1998). Imprecise terms in UK medical multiple-choice questions: What examiners think they mean. *Medical Education, 32*(4), 343-350.

Lowenthal, P., & Hodges, C. (2015). In search of quality: Using quality matters to analyze the quality of massive, open, online courses (MOOCs). *The International Review of Research in Open and Distributed Learning, 16*(5).

Margaryan, A., Bianco, M., & Littlejohn, A. (2014). Instructional quality of massive open online courses (MOOCs). *Computers & Education, 80*, 77-83. doi:10.1016/j.compedu.2014.08.005

Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010). The quality of a PeerWise MCQ repository. Paper presented at the *Proceedings of the Twelfth Australasian Conference on Computing Education-Volume 103,* 137-146.

Schuwirth, L. W., & 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.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*(8), 662-671.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. *Nurse Education Today, 30*(6), 539-543.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distracters in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*, 40-6920-9-40. doi:10.1186/1472-6920-9-40