Costello et al. International Journal of Educational Technology in Higher Education (2018) 15:42
https://doi.org/10.1186/s41239-018-0124-z

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

The future of online testing and assessment: question quality in MOOCs

Eamon Costello[1*], Jane Holland[2] and Colette Kirwan[1]

* Correspondence: eamon.costello@dcu.ie
[1]Dublin City University, Dublin, Ireland
Full list of author information is available at the end of the article

Abstract

If MOOCs are to play a role in the future of higher education it is imperative that we critically examine how they are currently functioning. In particular, questions persist about the role MOOCs will play in the future of formal accredited learning. As the focus turns from informal and free to formal, accredited and paid, greater scrutiny will be brought to bear on the quality of the courses themselves. Although there have been some empirical studies into the quality of MOOCs, a notable gap exists in that such research has not examined Multiple Choice Questions (MCQs) which are a key component of much MOOC assessment and testing. Previous research suggests that flawed MCQ items may compromise the reliability and validity of these assessments, potentially leading to inconsistent outcomes for students. This study was hence designed to examine MCQ quality in MOOCs. 204 MCQs were analysed, from a selection of 18 MOOCs, sampling the domains of computing, social science and health sciences. Over 50% of MCQs (112) contained at least one item flaw; 57 MCQs contained multiple flaws. A large proportion of MOOC MCQs violated item-writing guidelines, which is comparable with previous studies examining the prevalence of flaws in assessments in more traditional educational contexts. The problem of low quality MCQs can be ameliorated by appropriate faculty training and pre- and post-test quality checks. These activities are essential if MOOCs are to become a force that can enable enhanced and improved pedagogies in the future of higher education, instead of simply proceeding to replicate existing poor practices at scale.

Keywords: MOOCs, Educational futures, Tests, Quality, Multiple choice questions, Assessment

Introduction

Although some of the hope, hype and hysteria of the MOOC phenomenon may have abated, the number of courses running and of learners enrolling continues to climb (Class Central, 2018). The "gadarene rush" to MOOCs described by Daniel (2012), now looks less of a gold-rush or educational technology bubble. Despite their persistence, however, the final place that the MOOCs will claim in the future higher educational landscape is still uncertain. It may be that MOOCs find their best purpose as an "innovation platform", as a place for institutions to learn and experiment with pedagogy in an environment very different from their traditional institutional educational provision (Brown et al., 2015). Hence faculty who hold traditional pedagogical or teaching beliefs may be compelled to transform their practices through teaching

MOOCs (Freitas & Paredes, 2018). A recent study by Zhu, Sari and Lee (2018) of the methodological approaches to MOOC research, for instance, highlights the breadth of the research into teaching and learning at scale and the ambition to innovate traditional practice.

To achieve the greatest impact, MOOC pedagogical innovations must be brought home and mainstreamed. The affordances of MOOCs will need to be harnessed and married to the more traditional forms and mechanisms of higher education. To this end, there have been efforts to posit MOOCs as a bridge to more formal learning qualifications. This can potentially take many forms, one of which is designing systems through which MOOC learning can itself be accredited (Bralić & Divjak, 2018). Most of the major MOOC platforms such as edX, Coursera and FutureLearn have announced initiatives, promising to allow learners to attain accredited degrees through MOOCs (Baker et al., 2018). The Georgia Tech Masters in Computer Science through edX is one of the most prominent, and perhaps also successful, of these efforts to use MOOCs as a platform to deliver degrees. A study by Goodman et al. (2016) claimed that the MOOC had effectively opened new opportunities to gain formal degrees to learners who would otherwise not have gone to college, which is one of the few studies that give hard evidence that a MOOC has widened access to education in a significant way.

As the focus in the future turns from informal and free, to formal, accredited and paid - greater scrutiny will of course be brought to bear on the quality of the courses themselves. There have been empirical studies into the Quality of MOOCs overall (Lowenthal & Hodges, 2015; Margaryan et al., 2015), including the various assessment methodologies within (Sandeen, 2013; Balfour, 2013; Conole, 2016). Creating, and subsequently grading, valid assessments for courses with student cohorts numbering well into the thousands, is challenging. Individual grading by teachers or faculty becomes unfeasible, and so other approaches, such as automated marking and peer-assessment are employed, the latter relying heavily on learners' active participation (Admiraal et al., 2014; Balfour, 2013; Meek et al., 2017). However, a notable gap exists in that this prior research has not examined Multiple Choice Questions (MCQs), a key component of many MOOCs which employ automated assessments, in any way.

MCQs are omnipresent in education, and are a time-efficient method of assessment, enabling representative sampling of broad areas of course content(Van Der Vleuten & Schuwirth, 2005). They are utilised in most domains of study, but are particularly prevalent in disciplines such as medicine and the sciences, where an extensive body of literature now exists regarding their use (Schuwirth & van Der Vleuten, 2004). Moreover, mathematical models for assessing the reliability of this examination format, and item metrics such as difficulty and discrimination, are well-established (De Champlain, 2010). In addition to their use in summative examinations, tests (or quizzes) incorporating MCQs are also frequently found in formative assessments. They may be used in conjunction with Classroom Response Systems, or within innovative peer assessment systems such as PeerWise, which facilitates the development of student-generated learning tools, test-enhanced learning and peer assessment via MCQs (Denny et al., 2008; Larsen et al., 2008).

As with any assessment tool, there are potential criticisms and disadvantages regarding the MCQ format, particularly if used in isolation (Downing, 2002a; Schuwirth & van Der Vleuten, 2004). Simply-phrased questions may only test simple factual recall,

and be answered by simple repetition or recitation of memorised facts, perhaps with minimal actual understanding of content (Schuwirth & van Der Vleuten, 2004). This testing of factual recall may be an appropriate and valid component of assessment at some levels, but the ability to test application of knowledge is essential at higher cognitive levels. The incorporation of layered and detailed scenarios (contextual vignettes) within the stimulus format, followed by a 'lead-in' question designed to test the understanding or application of information thereof, allows testing of these higher cognitive levels and problem-solving abilities (Case & Swanson, 2002; Schuwirth & van Der Vleuten, 2004). In addition, it is essential that care is taken when writing multiple choice questions in order to minimise the possibility of *cueing*, the effects of which may be either positive or negative. Most of these flaws can be avoided with appropriate staff training and quality control procedures (Jozefowicz et al., 2002; Schuwirth et al., 1996; Tarrant et al., 2006a, 2006b).

While various formats have been employed since the introduction of MCQs over the past century, those most commonly seen in current use fall into two main categories, true/false or one-best-answer, which pose very different tasks for the examinee (Haladyna et al., 2002; Case & Swanson, 2002). Within the true/false format, the candidate is required to decide whether each option provided is true or false. This frequently means that they must also make a value judgment as to what *extent or degree* an option is correct; this may be straightforward, or may involve also trying to anticipate what the examiner had in mind when phrasing the question (Case & Swanson, 2002). In order to reduce potential ambiguity, item writers may default to assessing the recall of precise details and minutiae, instead of testing the understanding of broader, but perhaps more nuanced, concepts. In comparison, for one-best-answer MCQs, the candidate is simply required to rank the options *in order*, and then choose the best or most appropriate option from the list. Current guidelines from both North American medical licensing institutions with regard to writing MCQs advise the use of one-best-answer formats, for example either single best answer (SBA) or extended matching questions (EMQ) (Case & Swanson, 2002; Wood & Cole, 2001). The number of response options within each format is not entirely fixed, and institutions may vary this slightly according to their assessment strategy (Rodriguez, 2005; Swanson et al., 2008; Tarrant & Ware, 2010). That being said, most institutions will provide between 3 and 5 response options when writing SBAs, and 10 to 15 options within the EMQ format.

At a most basic level, MCQs may be required to undergo peer-review prior to use, or may be more formally assessed using a simple rubric (Jozefowicz et al., 2002); under certain circumstances, students themselves may be involved in this process and often prove to be discerning evaluators (Denny, Luxton-Reilly & Simon, 2009; Purchase et al., 2010). More comprehensive manuals and publications are also available to educators, advising on the design of appropriate assessment strategies, and offering recommendations and guidelines regarding the writing of high-quality MCQs. The canonical source is generally held to be that from the US National Board of Medical Examiners, but other published guidelines also offer excellent advice, in varying contexts (Case & Swanson, 2002; Tarrant et al., 2006a, 2006b; Wood & Cole, 2001; Haladyna et al., 2002). Despite the availability of these resources, the prevalence of flawed items in high-stakes examinations is an ongoing problem (Jozefowicz et al., 2002; Rodríguez-Díez et al., 2016; Tarrant et al., 2006a, 2006b). This is of concern due to indications that

that flawed MCQs may be confusing to examination candidates, penalizing some examinees (particularly non-native speakers), and thus potentially reducing the validity of the examination process (Downing, 2005; Tarrant et al., 2009). At their most egregious, item flaws may lead to MCQ assessments being guessable or gameable by 'test-wise' candidates, who otherwise have little content knowledge (Poundstone, 2014; Case & Swanson, 2002; Haladyna et al., 2002). Given the extent to which MCQs are relied upon to provide an accurate assessment of candidates' competence in both undergraduate and postgraduate medical education, there is a high demand for high-quality MCQs. Nonetheless, research over the past decade indicates a high prevalence of flawed items, in numerous contexts; this prevalence typically ranges from 33% to 72%, but with one publication reporting that *all* items reviewed contained at least one flaw (DiSantis et al., 2015; Rodríguez-Díez et al., 2016; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006a, 2006b; Downing, 2002b; Pais et al., 2016).

The question then arises as to the impact that flawed or ambiguous items may have on candidates. As previously stated, most authors group the potential effects into two main categories; those that introduce irrelevant difficulty into the question, and those that enable test-wise candidates to perform well. Within Table 1, we have categorized these flaws according to *where they appear within the item*, and outlined their potential effects as described within the evidence-base. These effects are not minor; 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, 2002b; Downing, 2005). However, interaction between flawed items and student achievement can be complex, and the effect of flawed items may not be consistent. Another publication examined MCQs used within undergraduate nursing examinations, and determined that 47.3% of all items were flawed (Tarrant & Ware, 2008). In contrast to Downing, 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). They also concluded that flawed items negatively impacted the high-achieving students in examinations, lowering their scores.

MCQs are currently a key element of MOOC assessments. The ease with which they can be deployed at scale, with automated marking, makes them particular well-suited for this medium. The results that participants receive from these assessments frequently contribute to their summative grade, and hence ultimately towards a certificate or credentials that they receive upon completion or participation, whether this be formal or informal in nature. Given this integral role, the question arises as to their quality. This issue is critically important if MOOCs are to fulfil aspirations to deliver formal learning that can contribute towards recognized awards. Therefore, in this study we sought to determine the prevalence of flawed items in a sample of MOOC MCQ assessments. Our study makes an important contribution by being the first empirical study to our knowledge that has critically examined MCQs in MOOCs.

In the remainder of this paper we will outline the methods of our study which include our how we collected the data that comprised our sample. We then describe how

Table 1 Framework of item flaws and their potential effect

Technical item flaws or unadvisable formats		Potential effect
Outmoded item formats		
True / false format		True-false questions require that examinees decide if a statement is true or false – at times a difficult decision, in a world where absolutes are rare. In addition, the examinee may also have to make a value judgment as to what *extent or degree* an option is correct; this may be straightforward, or may involve also trying to anticipate what the examiner had in mind when phrasing the question. (Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
Overly complex, or K-type, questions; e.g. choose option A if statements 1 and 2 are correct, choose option B if statements 1 and 3 are correct etc.		This format introduces unnecessary complexity to the format, increasing reading time, *construct irrelevant variance*, and reducing validity. (Case & Swanson, 2002, Downing, 2002a, 2002b, Jozefowicz et al., 2002, Haladyna et al., 2002, Tarrant et al., 2006a, 2006b)
Fill in the blank		These questions may be linguistically difficult to write, without giving grammatical clues to the examinee, and so are best avoided. (Downing, 2002a, 2002b)
Question ambiguity or obscurity		
Gratuitous information in stem		Inclusion of irrelevant information introduces unnecessary complexity to the format, increasing reading time, *construct irrelevant variance*, and reducing validity. Stems should be focussed, and only information relevant to answering the question should be included. (Case & Swanson, 2002, Downing, 2002a, 2002b, Tarrant et al., 2006a, 2006b)
Ambiguous or unclear information		Poorly worded questions can confuse examinees, even those of high ability, and are particularly problematic for non-native speakers. (Downing, 2002a, 2002b, Tarrant et al., 2006a, 2006b)
Unfocussed stem		Questions should be clear and explicit, with a definitive question (Haladyna et al., 2002, Tarrant et al., 2006a, 2006b)
Absolute terms		Elimination of options containing the words "always" and "never" greatly improves examinees' chances of choosing the correct option by chance. In addition, even supposedly absolute terms such as *"always"* or *"never"* may be interpreted differently, and means that examinees must make a value judgment as to what the writer means by the term in this context. (Holsgrove & Elzubeir, 1998, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
Vague frequency terms		Frequency terms are interpreted very differently by individuals, and their use means that examinees must make a value judgment as to what the writer means by a given frequency term in an individual question context. (Case, 1994, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
Negatively worded stem		Some writers advocate that negative stems *may* occasionally be used, so long as care is taken to phrase them simply and unambiguously. Others hold that high-quality negative MCQs are difficult to write well, and that their inclusion among otherwise positively-phrased questions may be confusing for examinees. (Haladyna et al., 2002, Jozefowicz et al., 2002, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
Structural or logical flaws		
Options & Stem	Problem is in the options not in the stem	The problem or question of the MCQ should be in the stem, not within the options. Inclusion of the problems within the options instead reduces the format to a true / false, or even a K-type complex format, with all the problems inherent within (above). (Haladyna et al., 2002, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
	Logical cues in stem & correct option	Logical cues (grammatical or numerical) in the stem and options may enable examinees to guess the correct option without any content knowledge. (Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
	Word repeats in stem & correct answer	Similar wording in the stem and options enables examinees to guess the correct option without any content knowledge. (Case &

Table 1 Framework of item flaws and their potential effect *(Continued)*

Technical item flaws or unadvisable formats		Potential effect
		Swanson, 2002, Downing, 2002a, 2002b, Tarrant et al., 2006a, 2006b)
Options	Longest option is correct	Writers often have an inherent bias to take extreme care in making the correct option with exact information and precise grammar, increasing the length; examinees may guess the correct option by assuming that the correct option is also the longest. (Case & Swanson, 2002, Downing, 2002a, 2002b, Tarrant et al., 2006a, 2006b)
	Implausible disctractors	Elimination of implausible distractors greatly improves examinees' chances of choosing the correct option by chance. (Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
	More than one, or no correct answer	A move away from the ***one-best-answer*** approach instead reduces the format to a true / false, or even a K-type complex format, with all the problems inherent within (above). (Haladyna et al., 2002, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
	Use of all of the above	The use of "all of the above" means that students may correctly eliminate this option by identifying at least one other response as being incorrect. (Haladyna et al., 2002, Tarrant et al., 2006a, 2006b)
	Use of none of the above	The use of "none of the above" means that students may correctly eliminate this option by identifying at least one other response as being correct. (Haladyna et al., 2002, Pachai et al., 2015, Case & Swanson, 2002, Tarrant et al., 2006a, 2006b)
	Option order / position of the correct option	Listing options in a consistent order on printed examination papers avoids bias on the part of the examiner towards edge aversion, i.e. a reluctance to place the correct option in the first or last positions. In a five option MCQ, this may result in option C being correct more often than would be expected by chance. Online examinations may be programmable to randomise option order, and so avoid this issue. (Case, 1994, Tarrant et al., 2006a, 2006b)
	Convergence cues	Writers often have an inherent bias to write distracters derived from the correct option, altering minor words or components; examinees may guess the correct option by choosing the one in which most option components appear together (Case & Swanson, 2002, Tarrant et al., 2006a, 2006b).

Technical flaws or unadvisable

we drew on relevant literature to develop our evaluation framework. The analysis of these results by the evaluators is then presented and followed by a discussion of the findings and their significance.

Methods

In the following section we detail how we set about the collection of the data that comprise this study. We detail how we drew on the relevant literature to develop framework for an analysis of that data and how this analysis was carried out both computationally and by the two human evaluators.

Sampling strategy and data collection

While we wished to sample MCQs from a broad range of courses, there were of necessity some specific inclusion criteria and practical limitations. We limited our selection to MOOCs delivered and assessed through the English language, and in areas where the authors had content knowledge and expertise; primarily the domains of computing,

social science and health sciences. Following an analysis of existing platforms, aggregators and previously published research, we identified approximately 300 courses as suitable for our purposes (Margaryan et al., 2015). As we needed to enrol in MOOCs and take the quizzes there were some practical limitations to the data collection which was labour intensive. Hence we sampled MOOCs that were running during the data collection period and this does not represent a true random sample but rather a convenience sample. The dataset published here represents a sample of 18 courses from the edX, Coursera, FutureLearn, Iversity and Eliademy MOOC platforms. This resulted in thirteen courses in the area of Computer Science, two each from Humanities and Health Sciences and one each from the domains of Psychology and Mathematics. Sixteen of these courses came from established higher educational institutions, one from a professional individual and another from a non-profit institute.

Data collection was performed manually; a total of 204 MCQs were collected for analysis from the selection of 18 MOOCs, a mean of 11.4 MCQs per MOOC. Each question and all of the options incorporated within, were copied from the MOOC quizzes and entered into a spreadsheet. The correct option was then identified in each case by the evaluators, and this information was recorded within the spreadsheet also. The evaluators had expertise in the areas of medical education and computer science (evaluator 1) and education and computer science (evaluator 2). The number of options per question ranged from 1 to 7, with a median of 4 options per MCQ observed.

Framework development

Our approach in evaluating the selected MCQs was to adopt a theory-driven analysis, using a priori themes drawn from extant research. Upon reviewing existing frameworks, we proceeded to adopt a template of 19 item-writing flaws (Table 1), primarily drawn from studies in the domain of health professions education (Tarrant et al., 2006a, 2006b; Haladyna et al., 2002). The use of this framework offered the potential for direct comparison between our results and those from previous work in alternative domains (DiSantis et al., 2015; Haladyna et al., 2002; Pais et al., 2016; Rodríguez-Díez et al., 2016; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006a, 2006b; Tarrant & Ware, 2008). The evaluators discussed the proposed framework, and then agreed on a common approach to the evaluation of specific flaws; decisions regarding ambiguous or nuanced item flaws were discussed and reached by consensus. Furthermore, we sought to improve our instrument by piloting a prototype on a smaller sample of data (114 MCQs) and seeking feedback from experts in the field at a conference (Costello, Brown & Holland, 2017).

Data analysis

With regard to data analysis, the item flaws included within our framework were identified in two ways; in a computational/automated way or manually. Automated evaluation was possible for five of the flaws incorporated within our framework: inclusion of options such as ‘all of the above’ or ‘none of the above’; the position of the correct option; the longest option being the correct one; and the inclusion of more than one, or no correct option. The length of each of the options was automatically calculated by counting the number of characters contained within it. The number of the options and the number of correct options were calculated in a similar manner, as was the position

of the correct option. The strings of 'all of the above' and 'none of the above' were programmatically detected.

The remaining thirteen items in our framework required manual coding by human evaluators according to the pre-defined framework, as agreed during original discussions. An initial subset of MCQs was assigned to two evaluators; both evaluators independently coded these assigned MCQs, then met to compare their evaluations and reach consensus. Inter-rater reliability was calculated and a Cohen's Kappa score of 0.92 was found which indicated a good degree of agreement. The evaluators then reviewed the remainder of the items, and met again to compare results. Once again, the evaluators discussed and reached consensus regarding the coding of MCQs within the dataset. Descriptive and inferential statistics were then performed on this data using the statistical software package R.

Results

In total, 204 questions were reviewed and a total of 202 item writing flaws were detected (Table 2). In 112 (54.9%) of the questions analysed there was at least one flawed item present, while 57 questions (27.9%) contained two or more flawed items. These were not evenly distributed among the MOOCs; when grouped by source, three MOOCs were found to have only a single flawed item among their sampled MCQs, but all the other courses had more than one flawed item in their sampled questions and every revised MOOC exhibited flawed MCQ items.

Table 2 Prevalence of item flaws in 204 MCQs from 18 MOOCs

Item Flaw	Number detected	% of MCQs
Outmoded item formats		
True / false format	22	10.8%
Overly complex, or K-type, questions	17	8.3%
Fill in the blank	4	2.0%
Question ambiguity or obscurity		
Gratuitous information in stem	0	0.0%
Ambiguous or unclear information	19	9.3%
Unfocussed stem	11	5.4%
Absolute terms	30	14.7%
Vague frequency terms	5	2.5%
Negatively worded stem	21	10.3%
Structural or logical flaws		
Problem is in the options, not in the stem	6	2.9%
Logical cues in stem & correct option	5	2.5%
Word repeats in stem & correct answer	8	3.9%
Longest option is correct	See inline text	
Implausible distractors	15	7.4%
More than one, or no correct answer	18	8.8%
Use of *"all of the above"*	2	1.0%
Use of *"none of the above"*	2	1.0%
Option order / position of the correct option	See inline text	
Convergence cues	9	4.4%

A large number of these flaws could be attributed to the use of older or outmoded formats; over 10% of questions (22 of 204) were written using the true/false format, and 17 were written in an overly complex manner (Tables 1 & 2). Similarly, a high number of items were either ambiguous or unfocussed in their phrasing of information provided ($n = 19$) or the question asked ($n = 11$), potentially leaving candidates unsure as to how to answer the question. Other flaws frequently observed were the inclusion of absolute terms ($n = 30$), negatively worded stems ($n = 21$), implausible distractors ($n = 15$) and the provision of more than one, or no correct answer ($n = 18$). Some flaws were infrequent, such as the inclusion of gratuitous or unnecessary information in the stem ($n = 0$), or the inclusion of the options 'all of the above' (n = 2) or 'none of the above' (n = 2).

In addition to the above flaws, the length of the options in each MCQ was calculated, by counting the number of characters contained within it, and we then examined how frequently the longest option was also the correct option, by means of Chi—squared analysis. The longest option was found to be correct significantly more often than would be expected by chance (χ^2 [1, $N = 204$] = 10.75, $p < 0.01$). With regard to calculating the position of the correct option, this was somewhat complicated by the fact that the number of options that may be provided with an MCQ is not fixed (Rodriguez, 2005; Swanson et al., 2008; Tarrant & Ware, 2010). Therefore, 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 113 MCQs to consider (Fig. 1). While a relatively even distribution was observed between the second, third and fourth positions, the first option was infrequently correct; however, these differences did not reach statistical significance (χ^2 [1, $N = 113$] = 7.46, $p = 0.06$).

Discussion

This descriptive study examined the prevalence or frequency of item writing flaws within MOOC MCQ assessments. While a high number of flawed items were identified, these results are comparable with previous research in other domains, although this is not perhaps a particularly reassuring fact (DiSantis et al., 2015; Haladyna et al., 2002; Pais et al., 2016; Rodríguez-Díez et al., 2016; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006a, 2006b;Tarrant & Ware, 2008; Downing, 2005). Additional studies examining the effect of flawed items on aspects of assessment such as item

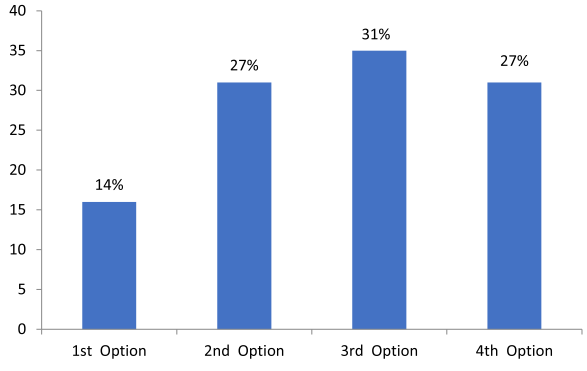

Fig. 1 Position of Correct Option (113 four-option MCQs)

psychometrics and candidate performance, suggest that these flawed items reduce the validity of the examination process, penalizing some examinees (Downing, 2005; Tarrant & Ware, 2008).

A large number of these flaws can be attributed to the use of older or outmoded formats; while some authors still hold a preference for the true/false format, these can be problematic, particularly if attempting to assess application of knowledge in context, as opposed to simple, explicit, factual recall (Case & Swanson, 2002). Likewise, overly complex item formats increase irrelevant item difficulty, and may disproportionally affect learners who are non-native speakers of English (Tarrant & Ware, 2008). If MOOCs are to realise ambitions of opening up access to education on a global scale assessment formats which disadvantage non-native speakers of the primary language of the MOOC present a problem.

A number of other item flaws were identified, which hold the potential to increase the *construct irrelevant variance*, or irrelevant difficulty, of the questions. These flaws may make the item harder to answer correctly, but are unrelated to the *content* purportedly being examined, and so are of little benefit in discriminating between candidates of high- or low-ability (Downing, 2002a; Rodríguez-Díez et al., 2016). For example, the use of a negatively worded stem was observed in 21 MCQs (10.3%) in our dataset. These items may often be quicker to construct, and so there may be reluctance from some writers to move away from this phraseology (Haladyna et al., 2002). Indeed, this is a commonly occurring item flaw, typically observed in up to 13% of MCQs in recent studies (Stagnaro-Green & Downing, 2006; Tarrant et al., 2006a, 2006b; Downing, 2005). However, students may find the wording of these items confusing, particularly if double-negatives or additional exemptions are introduced, and so this wording is discouraged in current item-writing guidelines (Tarrant et al., 2006a, 2006b; Case & Swanson, 2002; Haladyna et al., 2002; Wood & Cole, 2001).

One frequent criticism of the MCQ format is that the questions may be phrased in such a way as to be easily guessable to a 'test-wise' candidate (Poundstone, 2014). The inclusion of implausible distractors, seen in 15 MCQs within our dataset (7.4%), increases the possibility that a candidate may guess the correct option by random chance. The presence of logical cues or word repeats in both the stem and options may guide a candidate towards the correct option, even if they possess little or no intrinsic content knowledge (Case & Swanson, 2002; Tarrant et al., 2006a, 2006b). The most frequently occurring 'test-wise' flaw observed in our dataset was the presence of absolute terms, which was detected in 30 items (14.7% of MCQs; Table 2). While some hard sciences may lend themselves to absolutes, in general variance and individuality are intrinsic parts of life; question writers cannot always account for every possible circumstance ('never' might hold true today but not tomorrow). A further complication is that even supposedly absolute terms such as 'always' or 'never' may be interpreted differently (Holsgrove & Elzubeir, 1998). Test-wise candidates will be aware that the elimination of options containing the words 'always' and 'never' greatly improves their probability of choosing the correct option by chance. For these reasons, the use of absolute terms, in either the stem or options, is strongly discouraged (Case & Swanson, 2002; Holsgrove & Elzubeir, 1998; Tarrant et al., 2006a, 2006b). Best practice also recommends against using relative or absolute frequency terms (e.g. adverbs such as 'frequently','sometimes'), again due to the multiple interpretations that may be inferred (Case, 1994; Holsgrove &

Elzubeir, 1998; Tarrant et al., 2006a, 2006b). The potential for learners to be able to guess answers and become test-wise raises a question mark over the current fitness of these tests for formal learning. As such it represents a threat to the trustworthiness of MOOC accreditation which has the potential to undermine their future role in this regard.

It is beyond the scope of this study to say the MCQs analysed, even those functioning ones, contributed to anything beyond factual recall. Valuable lessons that can be drawn from this study are the clear importance of proper faculty training in MCQ writing, and the importance of quality assurance and review. Detailed guidelines with regard to the writing of high-quality MCQs do exist, and ample evidence now exists which shows the extensive, but often unpredictable effects which item flaws have on assessment psychometrics and student performance (Downing, 2005; Tarrant & Ware, 2008; Pais et al., 2016). 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. However, even some simple procedures could be included in MOOC MCQ engines in the future to obviate obvious flaws; the randomisation of option position is a simple matter when assessments are delivered online, but one which surprisingly few MOOCs seem to enforce. Many other common flaws could be detected through algorithmic means (Vasiliki et al., 2015). 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). Post-test statistical analysis of MCQs by established methods such as Classical Test Theory, would provide additional information regarding which items are poorly discriminating, and so potentially assist human evaluators in identifying item flaws. Our work here contributes by alerting researchers developing future testing and assessment in MOOCs to common pitfalls they may encounter.

The lack of these psychometric data was a limitation of this study. Hence, we are confined to a descriptive study, a simple representation of the prevalence of item flaws within a convenience sample of MOOCs. However, in the absence of any previous studies of this nature, this publication represents an opening of an important dialogue regarding the quality of these assessments, and the implications should they be used for formal accreditation into the future. The evidence from other domains is conclusive: item flaws are a threat to assessment validity (Downing, 2002a; Downing, 2005; Pais et al., 2016; Tarrant et al., 2006a, 2006b; Tarrant & Ware, 2008). If the MCQs from MOOCs analysed within our dataset were used in formal accredited assessments, it is credible that undesirable effects might occur on student achievement, with some students passing tests beyond their inherent ability, simply because of the presence of flawed items within the tests. It is vital that these assessments are fit-for-purpose, and we believe that rich bodies of research exist that can help define, develop and ensure quality in future online courses. All assessment methods have varying strengths and weaknesses, it is appropriate that multiple constructs and multiple data points are employed, particularly in respect of high-stakes decisions, if MOOCs are to be integrated within formal, accredited, degrees and courses (Sandeen, 2013; Epstein, 2007). So, MCQs may be best deployed as but one component of an overall assessment design that draws on several other existing assessment types such as self-assessment and peer

assessment particularly for formative assessment purposes (Admiraal et al., 2015) or that make use of semantic web technologies for more open ended questions (Del Mar Sánchez-Vera and Prendes-Espinosa, 2015). Balfour (2013) gives a useful analysis of the two assessment types comparing and contrasting semantic technologies, in this case automated essay grading, with calibrated human peer review in MOOCs. However, for all of the above approaches, authentication of identity, and potentially proctoring of examinations, remain challenges at the scales involved (Sandeen, 2013).

Conclusion

As the focus for the future of MOOCs is turning from informal and free, to formal, accredited and paid the need for critical appraisal of MOOCs has never been greater. This study aimed to address one specific aspect of MOOC assessments, the quality of MCQs, in order to better appraise where we are going and what future role MOOCs may best play. We found a high prevalence of errors in MCQs in the MOOCS we analysed which has potential detrimental effects on outcomes for learners. This may undermine trust in MOOCs themselves, particularly as vehicles for formal accredited learning. We recommend that greater training be given to staff engaged in developing MCQs for MOOCs and that staff be supported by peer review of their developed MCQs. We also highlight areas of future development for MOOC Platforms to help both check and improve the quality of MCQs. Finally we conclude that if MOOCs are to fulfil the promises they hold for the future of higher education there is more work to be done. The future may still be bright but we are not there yet.

Acknowledgements
The authors are grateful to Professor Michael O Leary the Prometric Chair in Assessment at Dublin City University for a critical review of the manuscript.

Availability of data and materials
An anonymised version of the full dataset and the findings of the evaluations is available on request from the authors.

Authors' contributions
All authors reviewed the research literature. The first author organized the review work and was main responsible for the development of the manuscript. The first and second authors lead the development of the framework. The third and first authors participated in the evaluations. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Dublin City University, Dublin, Ireland. [2]Royal College of Sugeons in Ireland, Dublin, Ireland.

Received: 21 June 2018 Accepted: 3 September 2018
Published online: 04 November 2018

References

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in massive open online courses. *Electronic Journal of E-learning*, *13*(4), 207–216.

Admiraal, W., Huisman, B., & Van De Ven, M. (2014). Self-and peer assessment in massive open online courses. *International Journal of Higher Education*, *3*, 119–128.

Costello, E., Brown, M., & Holland, J. (2016). What Questions are MOOCs asking? An Evidence-Based Investigation. *Proc. European Steakholder Summit (EMOOCS)*, Graz, 211-221.

Baker, R., Passmore, D. L., & Mulligan, B. M. (2018). Inclusivity instead of exclusivity: The role of MOOCs for college credit. In C. N. Stevenson (Ed.), *Enhancing education through open degree programs and prior learning assessment*, (pp. 109–127). Hershey: IGI Global.

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review™. *Research & Practice in Assessment*, *8*, 40–48.
Bralić, A., & Divjak, B. (2018). Integrating MOOCs in traditionally taught courses: Achieving learning outcomes with blended learning. *International Journal of Educational Technology in Higher Education*, *15*(1), 2 https://doi.org/10.1186/s41239-017-0085-7.
Brown, M., Costello, E., Donlon, E., & Giolla-Mhichil, M. N. (2015). A strategic response to MOOCs: How one European university is approaching the challenge. *The International Review of Research in Open and Distributed Learning*, *16*(6), 98–115.
Brunnquell, A., Degirmenci, Ü., 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. https://doi.org/10.1177/0163278710370459.
Case, S. M. (1994). The use of imprecise terms in examination questions: How frequent is frequently? *Academic Medicine*, *69*(10 Supplement), S4–S6.
Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences*, (3rd ed.,). Philadelphia, PA: National Board of Medical Examiners.
Class Central (2018). https://www.class-central.com/report/moocs-stats-and-trends-2017/. Accessed 23 Jan 2018.
Conole, G. (2016). *MOOCs as disruptive technologies: Strategies for enhancing the learner experience and quality of MOOCs*, (pp. 1–18). RED: Revista de Educacion a Distancia.
Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*, *3*, 18. https://doi.org/10.5334/2012-18 Available at: https://jime.open.ac.uk/articles/10.5334/2012-18/.
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. https://doi.org/10.1111/j.1365-2923.2009.03425.x.
Del Mar Sánchez-Vera, M., & Prendes-Espinosa, M. P. (2015). Beyond objective testing and peer assessment: Alternative ways of assessment in MOOCs. *International Journal of Educational Technology in Higher Education*, *12*(1), 119–130.
Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008). PeerWise: Students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research*, (pp. 51–58). New York: Association for Computing Machinery (ACM).
Denny, P., Luxton-Reilly, A., & Simon, B. (2009). Quality of student contributed questions using PeerWise. In Proceedings of the Eleventh Australasian Conference on Computing Education-Volume 95 (pp. 55–63). Australian Computer Society, Inc.
DiSantis, D. J., Ayoob, A. R., & Williams, L. E. (2015). Journal Club: Prevalence of flawed multiple-choice questions in continuing medical education activities of major radiology journals. *American Journal of Roentgenology*, *204*(4), 698–702. https://doi.org/10.2214/AJR.13.11963.
Downing, S. (2002a). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241.
Downing, S. M. (2002b). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10 Supplement), 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. https://doi.org/10.1007/s10459-004-4019-5.
Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine*, *356*, 387–396.
Freitas, A., & Paredes, J. (2018). Understanding the faculty perspectives influencing their innovative practices in MOOCs/SPOCs: A case study. *International Journal of Educational Technology in Higher Education*, *15*(1), 5 https://doi.org/10.1186/s41239-017-0086-6.
Goodman, J., Melkers, J., & Pallais, A. (2016). Can online delivery increase access to education? (NBER working paper 22754). National Bureau of Economic Research. Available at: https://research.hks.harvard.edu/publications/getFile.aspx?Id=1435.
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5.
Holsgrove, G., & Elzubeir, M. (1998). Imprecise terms in UK medical multiple-choice questions: What examiners think they mean. *Medical Education*, *32*(4), 343–350.
Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine: Journal of the Association of American Medical Colleges*, *77*(2), 156–161.
Larsen, D. P., Butler, A. C., & Roediger 3rd, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*(10), 959–966. https://doi.org/10.1111/j.1365-2923.2008.03124.x.
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), 83–101 Available at: http://www.irrodl.org/index.php/irrodl/article/view/2348/3411.
Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education*, *80*, 77–83. https://doi.org/10.1016/j.compedu.2014.08.005.
Meek, S. E. M., Blakemore, L., & Marks, L. (2017). Is peer review an appropriate form of assessment in a MOOC? Student participation and performance in formative peer review. *Assessment & Evaluation in Higher Education*, *42*, 1000–1013.
Pachai, M. V., Dibattista, D., & Kim, J. A. (2015). A systematic assessment of 'none of the Above'on multiple choice tests in a first year psychology classroom. *The Canadian Journal for the Scholarship of Teaching and Learning*, *6*, 2.
Pais, J., Silva, A., Guimarães, B., Povo, A., Coelho, E., Silva-Pereira, F., … Severo, M. (2016). Do item-writing flaws reduce examinations psychometric quality? *BMC Research Notes*, *9*(1), 399. https://doi.org/10.1186/s13104-016-2202-4.
Poundstone, W. (2014). *Rock breaks scissors: A practical guide to outguessing and outwitting almost everybody*. New York, Boston and London: Little, Brown and Company.
Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010). The quality of a PeerWise MCQ repository. In *Proceedings of the Twelfth Australasian Conference on Computing Education, 103*, (pp. 137–146). Darlinghurst, Australia: Association for Computing Machinery (ACM).
Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x.
Rodríguez-Díez, M. C., Alegre, M., Díez, N., Arbea, L., & Ferrer, M. (2016). Technical flaws in multiple-choice questions in the access exam to medical specialties ("examen MIR") in Spain (2009–2013). *BMC Medical Education*, *16*(47), 1–8. https://doi.org/10.1186/s12909-016-0559-7.
Sandeen, C. (2013). Assessment's place in the new MOOC world. *Research & practice in assessment*, *8*, 5–12.

Schuwirth, L. W., Der Vleuten, C. V. D., & Donkers, H. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education*, *30*(1), 44–49.
Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974–979. https://doi.org/10.1111/j.1365-2929.2004.01916.x.
Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England journal of medicine for continuing medical education. *Medical Teacher*, *28*(6), 566–568. https://doi.org/10.1080/01421590600711153.
Swanson, D. B., Holtzman, K. Z., & Allbee, K. (2008). Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine: Journal of the Association of American Medical Colleges*, *83*(10 Supplement), S21–S24. https://doi.org/10.1097/ACM.0b013e318183e5bb.
Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006a). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *26*(8), 662–671. https://doi.org/10.1016/j.nedt.2006.07.006.
Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006b). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*, 354–363.
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. https://doi.org/10.1111/j.1365-2923.2007.02957.x.
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. https://doi.org/10.1016/j.nedt.2009.11.002.
Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*, 40. https://doi.org/10.1186/1472-6920-9-40.
Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, *39*(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x.
Vasiliki, G., Filippos, F., Christina, R., & Serafim, N. (2015). Software-assisted identification and improvement of suboptimal multiple choice questions for medical student examination. *Health Science Journal*, *9*(2), 8.
Wood, T. & Cole, G. (2001). Developing multiple choice questions for the RCPSC certification examinations. The Royal College of Physicians and Surgeons of Canada, Office of Education. Available at: https://www.macpeds.com/documents/GuidelinesforDevelopmentMCQRoyalCollege.pdf.
Zhu, M., Sari, A., & Lee, M. M. (2018). A systematic review of research methods and topics of the empirical MOOC literature (2014–2016). *The Internet and Higher Education*, 37, 31–39.