MULTIPLE CHOICE QUESTIONS: ANSWERING CORRECTLY AND KNOWING THE ANSWER

Peter McKenna
Manchester Metropolitan University, John Dalton Building, Manchester M1 5GD, UK

ABSTRACT

Multiple Choice Questions come with the correct answer. Examinees have various reasons for selecting their answer, other than knowing it to be correct. Yet MCQs are common as summative assessments in the education of Computer Science and Information Systems students.
To what extent can MCQs be answered correctly without knowing the answer; and can alternatives such as constructed response questions offer more reliable assessment while maintaining objectivity and automation?
This study sought to establish whether MCQs can be relied upon to assess knowledge and understanding. It presents a critical review of existing research on MCQs, then reports on an experimental study in which two objective tests were set for an introductory undergraduate course on bitmap graphics: one using MCQs, the other constructed responses, to establish whether and to what extent MCQs can be answered correctly without knowing the answer.
Even though the experiment design meant that students had more learning opportunity prior to taking the constructed response test, student marks were higher in the MCQ test, and most students who excelled in the MCQ test did not do so in the constructed response test. The study concludes that students who selected the correct answer from a list of four options, did not necessarily know the correct answer.
While not all subjects lend themselves to objectively testable constructive response questions, the study further indicates that MCQs by definition can overestimate student understanding. It concludes that while MCQs have a role in formative assessment, they should not be used in summative assessments.

KEYWORDS

MCQs, Objective Testing, Constructed-Response Questions

1. INTRODUCTION

Multiple Choice Questions (MCQs) are a well-known instrument for summative assessment in education: they typically require students to select a correct answer from a list of alternatives. Most typically, there will be a single correct answer among two, three or four options; though variations can include selection of a single best-possible answer, or of multiple possible answers ('multiple response').

MCQs are widely used as an assessment tool in education. Just how widely, and in what contexts, cannot be ascertained with any reliable degree of accuracy. Faris et al (2010) assert that they are "the most frequently used type of assessment worldwide." Bjork et al (2015) describe them as 'ubiquitous'. While they are not as useful in humanities subjects, MCQs are commonly deployed in several STEM subjects – including Computer Science - and by Professional, Statutory and Regulatory Bodies including those in critical areas such as health, pharmacy, law economics and accountancy. As they can be marked automatically – and, in principle, objectively – they will normally save staff time in terms of marking, moderation, and providing feedback.

It may be for this reason that the intrinsic pedagogic quality of a format that presents students with the answer is seldom questioned or tested. The use of MCQs is often accompanied by at least a perception of partisanship for or against them. Those who challenge MCQs as a reliable assessment tool can leave themselves open to accusations of bias and prejudice (Moore, 2104).

Literature on MCQs generally accepts their ubiquity and prioritises practical treatments: guidelines for optimising and construction (Dell and Wantuch 2017; Consodine et al 2005; Haladyna 2004; Bull and McKenna 2004; Morrison and Free 2001); ways of easing construction (Dehnad et al. 2014); and strategies for minimising the scope for guessing beyond the base mathematical probabilities (Bush 2015; Ibbot and Wheldon 2016).

The relative merits of different formats is well-examined: for example, Vegada et al (2016) found no significant performance difference between 3-option, 4-option and 5-option questions – and recommended using three. Dehnad et al (2014a) on the other hand found a significant difference between 3-option (better) and 4-option questions, but also recommended 3-options as easier for new teachers and easier to cover more content by saving question development time. They also suggest that 3-option questions are more reliable, in that having to provide four options would force teachers "to use implausible and defective distracters". There is also a significant body of literature investigating variations on the choice process such as subset selection testing, negative marking, partial credit, and permutational multiple choice. This paper will focus on the use of standard MCQs, where there is one correct answer among three, four or five options.

The popularity and status of MCQs appears to arise at least in part from the ease and efficiency with which technology – from optical mark scanners to JavaScript-enabled web environments - can produce results, particularly for large numbers of examinees. The adoption of MCQs can be seen as a "pragmatic" strategy (Benvenuti 2010) in response to large class sizes. Students also believe that MCQ tests as easier to take (Chan and Kennedy 2002); and McElvaney's (2010) literature review concludes that MCQ tests are not only common in universities but also "well accepted by students and teachers". Srivastava et al (2004) are unusual in presenting a position paper asserting that medical and surgical disciplines do not need students who can memorise information; that there is no correlation between such recall and clinical competences; and proposing that MCQ's be abolished from medical examinations and replaced with free response or short answer questions.

In 2014 Central Queensland University in Australia banned MCQs on the basis that they test a combination of guessing and knowledge, lack authenticity, misled learners with distractors, and were akin to game shows (Hinchliffe 2014). A paper subsequently written by academic staff at Western Sydney University (Ibbett and Wheldon 2016) cited "efficiency benefits" in defence of MCQs, but found that almost two-thirds of MCQs found in six test banks of cash flow questions, provided some sort of clue to the correct answer. Ibbett and Wheldon present the ways in which guessing could be minimized by improving the quality of questions and eliminating clues as proof of their potential 'reliability' and as a case against the 'extreme' measure of forbidding their use. They note past anticipation that cluing problems would be eliminated from test banks, and that in 2016 such aspirations were far from being fulfilled. While recognising the extent of the cluing problem in test banks, they did not appear to recognise any base level statistical guessability inherent in choosing a single correct answer from a small number of options.

The literature that deals with guessability largely focuses on good question design (Haladyna 2004); different uses (Nicol 2007; Fellenz 2010); debates concerning counteractive measures such as negative marking (Espinosa and Gardeazabal 2010; Lesage et al 2013); or reducing the basic odds from number-of-options to one via permutational multi-answer questions (Bush 1999; Kastner and Stangl 2011) and extended matching items (George 2003). Harper (2002) suggests that extended matching questions have "a detrimental effect on student performance" and that it may therefore be "safer" to use MCQs. The desire for efficiency can sometimes seem to occasion an element of misdirection: Boud and Felleti (2013) see MCQs as "the best way to assess knowledge gleaned from a [problem-based learning] experience" on the basis that short-answer questions do not measure anything distinctive in terms of problem-based learning. It is however illogical to equate the proposition that such questions do not measure anything distinct, with validity and reliability – as if this lack of distinction in the attributes to be tested extended to the results of any such testing.

This study examines whether MCQs can be answered correctly without knowing the answer. The literature on MCQs is considered, followed by a report on a test of the reliability of MCQ results when compared to short constructed responses in an area of Computer Science.

2. THE NATURE OF MCQS

2.1 The Numbers Game

The fact that MCQs present the correct answer, with the odds good for guessing which one it is, may be something of an elephant in the exam room. The per-question odds of 4 to 1 for standard one-correct-answer

out of four questions may be mathematically extended to test level, where a student who knows a third of the answers to thirty questions, will on average guess five out of the remaining twenty questions and thereby pass with a test grade of 50%. Where the pass mark is 40%, it would on average be necessary only to know six - one fifth - of the 30 answers: it is necessary then to guess correctly only a further six of the remaining 24 questions; and the probability of successfully guessing at least six is around 58%. There is a 5% probability of a student who knows nothing getting at least 12 questions right: five in every hundred students who know nothing will on average pass the test. Such odds assume optimally-written MCQs, with no clues or weak distractors: the reality is very often different, with studies that examined test banks for nursing and accounting education (Masters et al 2001; Tarrant et al 2006; Ibbett and Wheldon 2016) finding multiple problems in question formulation and quality and recurrent violations of item writing guidelines.

2.2 Using Flaws

While the problem of guessing is often ignored or deprioritised, it has also been reframed as something that is potentially useful: Bachman and Palmer (1996) suggest that informed (rather than random) guessing should not only be taken into account but actively encouraged, on the basis that it demonstrates "partial knowledge of the subject matter". In terms of question quality, Kerkman and Johnson (2014) have even turned poorly-worded MCQs into a learning opportunity enabling students to be rewarded if they challenge or critique questions.

Another issue identifiable with MCQs is the presentation of incorrect but plausible answers. In a series of tests, McDermott (2006) reports the "false recognition of related lures". As early as 1926, Remmers and Remmers reported on what they called "the negative suggestion effect" in true-false examination questions. McClusky (1934) noted that ability to recognise a false statement did not entail an equal ability to make it true. Roedeger and Marsh (2005) conclude that multiple choice testing can "create false knowledge or beliefs in students that they take away from the classroom. In domains such as language learning (where MCQs are also particularly deficient in authenticity) false models can present an approximation that may appear correct, while the correct form is not sufficiently embedded. This may also be reasonably said in the context of programming languages and algorithms.

2.3 What MCQs Test

Srivastava et al (2004) suggest that MCQs emphasise "recall of factual information rather than conceptual understanding and integration of concepts". Wainer and Thissen (1993) suggest that MCQs "may emphasise recall rather than generation of answers". (Dufresne et al. 2002) in the context of a Physics test concluded that "a correct answer on the chosen MCQ is, more often than not, a false indicator of deep conceptual understanding". Simkin and Kuechler (2005) conclude however that MCQs are not homogenous, and can – with greater difficulty - potentially test higher levels of understanding.

Just as recognition is easier than recall in terms of computer interface design (Johnson 2014) – epitomised by the difference between command-line and menu-driven interfaces - facts and concepts can more readily be recalled, and procedures recognised, if they are presented to the student. Fundamentally, MCQs provide examinees with the answer: the only challenge is to pick it out from the 'menu' of options. However, alternatives to MCQs are available that share much of their convenience and efficiency of scale, but do not provide the answer. Questions that require students to enter the answer, can range from fill-in-the-blank questions to short-essay questions. The former may also be susceptible to guessing, and the latter entails subjective scoring and cannot be meaningfully automated. (Wainer and Thissen 1993) report that a Chemistry test cost some 3000 times more than a comparable MCQ exam. This, however, assumes that subjective scoring is necessary.

It is nonetheless possible in some areas to test knowledge and understanding via the use of short constructed response questions (CRQs) or calculated questions that are simple, single-stage and not open-ended; can be automatically marked; and carry little or no scope for guessing. This is particularly the case where numerical answers can be calculated, based on a conceptual understanding and application of the principles and processes underpinning the calculation. In other fields subjectivity of marking is seen as a disadvantageous aspect of constructed response questions (McElvaney et al. 2012). Simkin and Kuechler (2005) list what they see as advantages of MCQ tests over constructed response tests –largely on the basis of

an assumption that the latter are not machine gradable and entail some subjectivity (and hence instructor bias) – and conclude that the perform "an adequate job" of evaluating student understanding. Others have asserted on the same basis that MCQ reliability is higher (Wainer and Thissen 1993; Kennedy and Walstad 1997). However, constructed response questions in some disciplines do not involve subjectivity and do still carry the same functional benefits as MCQs in terms of ease, consistency, speed and accuracy of marking.

The 2017 Australian Mathematics Competition included, in addition to twenty-five traditional MCQs, five higher-value questions that required an answer within the integer range 0-999. These were entered by means of pencil marks on a mark sense sheet, using three columns for place values with 10 rows for each representing numbers between 0 and 9 (Australian Mathematics Trust 2017).

Matters and Burnett (1999) found that omit rates were significantly higher for short-response questions than for MCQs. This may be hardly surprising, but it suggests that guessing does occur with the latter.

3. METHODOLOGY

3.1 Two Different Tests, Same Group

Two formative assessment tests were devised: one consisting of constructed responses, the other of one-correct-answer out of four multiple choice questions. The constructed-response test questions were formulated so that answers could be marked objectively. As long as the terms of the question were unambiguous, and/or any potential variations of the correct answer were permitted as answers, they could be marked both objectively and automatically.

Both tests were administered via Moodle, to a cohort of 280 students taking a Level 4 (first year undergraduate) multimedia unit. It was taken on an open book basis, and as formative assessment: none of the answers could be directly found by searching the Internet. As the students were first years, control on the basis of prior knowledge or ability was problematic. It was therefore decided to deliver both tests to all students. Clearly this could not be done simultaneously.

In an early study, Traub and Fisher (1977) used two identical tests, administering a free-response version two weeks before a multiple-choice version. They chose this order on the basis that doing so would eliminate learning from the cues found in the MCQs. (Like Boud and Felleti (2013), their focus was on equivalence of attributes tested rather than of results; and the marking of free-response answers was assumed to require an objectification process).

Based on the statistical potential for guessing the correct answer of an MCQ, the hypothesis was that students would score better in the MCQ test where the correct answer could be selected from a list of four, when compared to the equivalent CRQ test where the correct answer had to be typed into a field. If Traub and Fisher's sequencing were to be followed, with the CRQ test preceding the MCQ test, the potential to perform better in the latter – having already prepared for, taken, and reflected on a test - would have been enhanced. To counteract any bias towards the hypothesis, it was therefore decided to deliver the MCQ test first; and to allow a week between the tests. This introduced a bias towards better performance in the CRQ test, as students had an extra week to learn (including from the experience of taking the MCQs) and were taking the second test at a time when the topic might reasonably still be fresh in the mind. The MCQ test results were released after all students had sat it; but would be hidden during the CRQ test.

Constructed response questions were formulated so that the range of potential answers was large enough to eliminate guessing.

3.2 The Questions

In order to establish whether students performed better in an MCQ test compared to a similar CRQ test, two equivalent tests, consisting of MCQs, and CRQs respectively - were devised for a topic within a first year unit introducing bitmap graphics concepts. The topics chosen are not high-order learning, but they do test conceptual understanding and practical application of principles and techniques. Both sets of questions covered the same topics:

a) Identify how many colours can be represented by a given colour depth.

b) Identify the file size of an uncompressed 8-bit colour image of given pixel dimensions with a palette of a given number of colours.

c) Identify a colour from given RGB values

d) Identify the physical measurements of an image of given pixel dimensions on a monitor with a given resolution

e) For a given convolution mask applied to a given 24 bit RGB pixel value with a given set of neighbouring pixels, identify the new RGB value of the processed pixel

Question (a) involves applying a rule rather than recalling a memorised answer. Both the MCQ and the short-answer questions used an atypical colour depth (10 bit colour for the MCQ, 12 bit colour for the CRQ) that would not be susceptible to recollection. The atypical colour depth was chosen to eliminate partial knowledge, such as remembering rather than calculating the number of colours available with commonly used colour depths such as 8 and 24 bit. Possible answers for a CRQ would in theory include all positive integers.

Question (b) involves understanding of colour depth but is a more intricate calculation, based on further understanding of both bitmapping and colour lookup tables. The range of possible answers is in theory any positive integer. The MCQ asked students to choose the correct uncompressed file size of a 100x120 pixel 8 bit colour image with a 128 colour palette; the CRQ asked for the size in kilobytes to two decimal places of a 100x100 pixel 8 bit colour image with a 64 colour palette. To guard against guessing in the MCQ, the distractors were kept within a narrow range up to 4KB distant from the correct answer.

Question (c) is more constrained. Identifying a full range of named colours is neither intuitive nor necessary, so the question was limited to testing understanding of colour channel balance by identifying shades of greyscale. For the CRQ various permutations of grey (and gray) had to be provided as correct answers in the Moodle question editor. It is possible with this particular question that the previous week's multiple choice question may have provided clueing; and likely that a web search would yield the correct answer.

Question (d) tests understanding of resolution: the student is given the physical dimensions of a paper-based image along with the scan resolution and asked to identify its physical width in inches when displayed on a monitor with a given physical resolution. The answer is not restricted to the monitor dimensions and could potentially be any integer. The MC question distractors therefore occupied a wide range around the correct answer. The CR question instructed students to enter one integer only, but allowance was provided in the marking for variations in presentation, including suffixes to denote inches.

Question (e) assesses understanding and application of the method whereby image processing filters calculate new pixel values using convolution masks.

Given a specific convolution mask and specific pixel values for a pixel to be processed and for its neighbours, the new value for the processed pixel will consist of a combination of three colour channel values, each of which can have an integer value within the range 0 to 255. There are therefore 16,777,216 different possible answers that are legal. The odds of successfully guessing the correct answer in a CRQ are consequently higher than the odds of winning the UK lottery (RWAP Services, nd) or of being killed by lightning (Roper 2008).

In order to simplify short-answers entry and the requirements for parsing those answers, the convolution mask questions were designed to ask only for the value of a single given channel: for example, the new value of the green channel for the pixel to be transformed. The odds for successful guessing are thereby reduced from 2563 to 1, to 256 to 1 for a short-answer; they remain at 4 to 1 for an MCQ. Both questions were presented using a visual illustration of the convolution mask and the colour channel values (see Figures 1 and 2).

MCQs were written to ensure no clues or cues were present, and to provide credible distractors. No negative marking would be applied to the MCQ test – initially and at least for the purposes of clear results and feedback.

For all questions the corresponding MC and CR questions were very similar, but not identical, to ensure an equal level of difficulty but at the same time avoid any potential for carrying forward MCQ answers to the constructed response test. For example, the convolution mask MCQ and CRQ used the same mask, but different starting values:

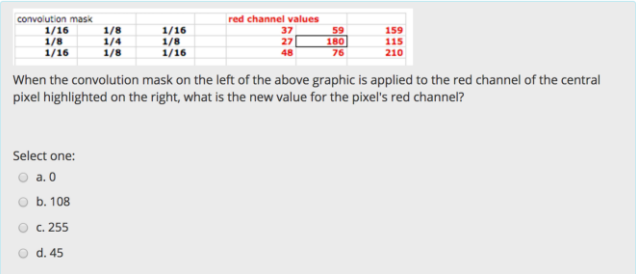

Figure 1. Sample Multiple Choice Question on Convolution Masks

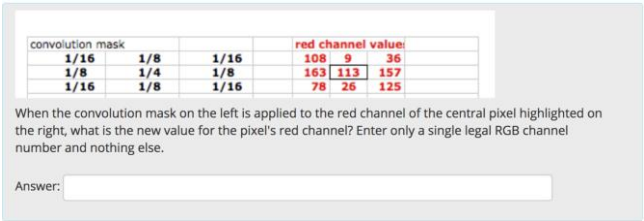

Figure 2. Sample Constructed Response Question on Convolution Masks

Of the 280 students in the cohort, 136 answered the MCQ test, and 105 answered the short-answer test. Of these, 78 attempted both. Of the 78 attempts, 14 did not complete and submit one or the other.

4. RESULTS AND DISCUSSION

4.1 Grades

Of the 64 students who took both tests, 32 obtained a better overall mark in the MCQ version, 11 obtained a better overall mark in the constructed response version, and 21 obtained the same mark in both.

The time taken to answer could not be reliably measured, as the time-taken data provided by Moodle showed that several students (three on the constructed responses quiz, six on the MCQs) left their attempt open for days before finishing.

Given the advantage of having previously taken the MCQ test and an additional week to study, the fact that half of the students did better in the earlier (MCQ) test (with two thirds of the others doing as well) might suggest that several students acquired marks for questions that they did not know the answers to.

More students (136) took the MCQ test than the CRQ (105). While this could conceivably be because it was the first of the two to be delivered, it might alternatively support the suggestion that MCQs are a more attractive option to students. Of the 105 who opened the CRQ test, almost a quarter - 25 - left it open and did not submit. Of the 136 who took the MCQ test, approximately 13% - 18 - left it open. One student scored zero in the CRQ test – the same student had scored 2 in the MCQ test a week earlier. Of the three students who scored zero in the MCQ test, two did not take the CRQ test and the third scored two.

Thirteen students scored full marks in the MCQ test; of these, only two also scored full marks in the CRQ test – with four scoring only 2 out of 6, three scoring 5 and one scoring 3. The remaining three did not engage with the CRQ test; all three however did engage with the final summative test five weeks later, with one scoring 30%, one 50%, and one 80% in the 10 CRQ questions on that test.

The results indicate that some individual students who successfully select the correct answer from a list of four alternatives, were able to do so without knowing the answer. As the MCQs were robustly designed to eliminate cues and clues, and the CRQs required students to know and understand the answer, this would suggest that guessing is a factor in the success of some students in MCQ tests.

4.2 Negative Marking

Negative marking is the most prominent formula scoring alternative to number right scoring. This scoring method aims to reduce or eliminate guessing by penalizing incorrect answers. Variations include awarding marks for unanswered items, as a further measure to discourage guessing (Prieto and Delgado, 1999; Campbell, 2015). Burton (2002) concluded that the impact of guessing with number right scoring was greater than that of variable attitudes to risk with negative marking. On the other hand, Bar-Hillel et al (2005) believe number right scoring to be superior to negative marking. – though this is an opinion based on the relative difficulty of communicating the latter properly, as well as the presence of various types of cues. They also suggest that 'tacit collusion' exists among all stakeholders based on intuitive but irrational expectations of marking schemes.

The most common formula for applying negative marking to incorrect answers is $1/(n - 1)$, where n is the number of choices. When a 0.33 penalty was applied to the MCQ test results, slightly more students who did both the MCQ and CRQ tests still obtained a higher than a lower result (39 against 35) in the MCQ compared to the constructed responses.

Assuming a pass mark of 40%, 45 students passed the MCQ test under the negative formula marking: approximately the same as those who passed the CRQ test (46), but substantially lower than those who passed with number right scoring (62).

While the output from the negative marking formula was much closer to the results achieved from the CRQ test, it was not possible to determine the extent to which the negative marking counteracted guessing; or to elicit any risk-taking variations.

Standard setting offers an alternative remedy: to raise the pass grade for number right scoring, based on the odds of correctly guessing answers. This is common in high-stakes, safety-critical areas: for example, the European Aviation Safety Agency sets a 75% cut score, with no negative marking, for its theoretical knowledge MCQ examinations (CAA, 2018). With only six questions in each test, however, testing this strategy would have lacked significance. Given the nature of the questions, it would be difficult to ensure a good level of student engagement with formative tests consisting of a large number of questions.

4.3 Partial Understanding

In tests with more open formats, it is it is possible to demonstrate partial understanding – and to receive marks for knowing the relevant parts of the answer. The same can even be true of CRQ tests: in the case of our constructed response questions, a student might for example understand the formula for calculating a transformed pixel value, but enter the incorrect answer due to an arithmetic error. In the MCQ version, the student might be able to identify their error and revisit the calculation accordingly. On the one hand, the question does not assess arithmetic; though on the other, it is a concomitant skill if the formula is to be applied. Answers are precise, and the value of getting 'close' is not necessarily significant. Providing an additional field for the entry of the formula may perhaps be a more practical solution than providing fields for calculation steps.

The value of partial understanding can vary greatly between and even within disciplines, and is not easy to assess objectively. In some cases where an answer is verbal, for example, we may know the answer when we see it: it can be recognised but not recalled. Partly-correct distractors will be selected by students who understand the correct part, as well as by those who guess. While it may be argued that plausible or partly-correct distractors can mislead students, implausible distractors are also bad practice in that they make guessing much easier. Polytomous scoring, whereby partial credit is given for partly correct distractors, may be useful where grading is norm-based (Grunert et al, 2013) and a sufficient number of options are provided to minimise guessing.

While single answer questions tend not to provide scope for the demonstration of partial understanding – unless negative marking is introduced (Bond 2013), or one considers 'informed' guessing to be both detectable and indicative of partial understanding (Bachman and Palmer 1996) - CR question (e) did afford a potential scope for incorrect answers that demonstrated partial understanding: were an unrounded raw number (outside of the 0-255 range) entered as the result it would, if it were correct as an unrounded number, show understanding of the process whereby the new pixel value is obtained. The missing understanding is, however, more basic than the understanding of the calculations that would produce that

unrounded number: that the range of rgb values is constrained to the 0-255 range (the question specified that the number should be a legal value for an RGB channel), and that calculated figures therefore have to be rounded up or down into that range. Such an instance of partial understanding would therefore seem very unlikely to occur, and the question is clear that only legal rgb values should be entered. However, this could be given partial credit if desired in both formats – though with a traditional four-answer MCQ providing the unrounded value as an option would increase the potential for productive guessing to at least 50:50.

5. CONCLUSION

The experimental study confirmed something that might reasonably be said to be obvious but which is often ignored: that students can answer MCQs successfully without knowing the answer. It also suggests that constructed response questions are a more reliable means of assessing understanding and knowledge. This again might seem obvious – but given the prevalence of MCQs in summative assessments, it is worth stressing. The subject area in the study lent itself well to constructed response questions that can be objectively tested: similar questions can be readily devised not just for calculated answers but also for restricted syntaxes such as programming languages. Even where answers are potentially less unambiguous, verbal variations and pattern matching can be used to cater for possible answers.

The same cannot be said for all subjects – within Computer Science or elsewhere. However, the principle is universally applicable to the use of MCQs in summative assessments. Well-designed MCQs can be very useful within the learning process by providing instant formative feedback. However, their role in summative assessments is problematic: most significantly because number right scoring overestimates student performance. The robustness of methods used to calculate (higher) cut scores fell outside the scope of this paper; in the absence of any such demonstrably effective method, number right scored MCQs should not be used in summative assessments. While negative marking produced broadly similar results, further research could usefully set MCQ and CRQ tests at the same time so that the accuracy of such formula scoring could be examined at an individual level.

REFERENCES

Australian Mathematics Trust. 2017. "Australian Mathematics Competition". 27 July 2017. Accessed July 28. http://www.amt.edu.au/mathematics/amc/amc-scoring-system/

Bailey, P.H.J. et al, 2012. Implications of multiple-choice testing in nursing education. *Nurse Education Today* 32:40–44.

Bar-Hillel, M. et al, 2005. Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society 4: 3. https://doi.org/10.1007/s11299-005-0001-z*

Benvenuti, S., 2010. Using MCQ-based assessment to achieve validity, reliability and manageability in introductory level large class assessment. *HE Monitor* 10 Teaching and learning beyond formal access: Assessment through the looking glass: 21-34.

Bjork, E.L. et al, 2015. Can Multiple-Choice Testing Induce Desirable Difficulties? Evidence from the Laboratory and the Classroom. *The American Journal of Psychology* 128, no. 2:229-239.

Bond, A.E. et al, 2013. Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety. *PLOS* ONE, 8, 2

Boud, D. and Feletti, G., 2013. *The Challenge of Problem-Based Learning.* Routledge.

Bull, J. and McKenna, C., 2004. *Blueprint for Computer-assisted Assessment*. London: Routledge Falmer.

Burton, R.F., 2002. Misinformation, partial knowledge and guessing in true/false tests. *Medical Education* 36 (2002), pp. 805-811

Bush, M., 1999. Alternative marking schemes for on-line multiple choice tests. In *7th Annual Conference on the Teaching of Computing*, Belfast.

Bush, M., 2015. Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education* 40: 2, 218-231, DOI: 10.1080/02602938.2014.902192.

CAA, 2018. *Theoretical knowledge examinations*. Available at: https://www.caa.co.uk/General-aviation/Pilot-licences/EASA-requirements/General/Theoretical-knowledge-examinations/ [Accessed 13 Apr. 2018].

Campbell, M.L., 2015. Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed To Discourage Guessing. *Journal of Chemical Education* 92 (7), 1194-1200. DOI: 10.1021/ed500465q

Chan, N. and Kennedy, P.E., 2002. Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and 'Equivalent' Constructed-Response Exam Questions. *Southern Economic Journal* 68: 957–971.

Considine, J. et al, 2005. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian* 12:1:19-24.

Dehnad, A. et al, 2014. Comparison between Three-and Four-Option Multiple Choice Questions. Proceedings of the International Conference on Current Trends in ELTA. *Procedia - Social and Behavioral Sciences* 98, 398–403.

Dell, K.A. and Wantuch, G.A., 2017. How-to-guide for writing multiple choice questions for the pharmacy instructor. *Currents in Pharmacy Teaching and Learning* 9.1:137-144.

Dufresne, R.J. et al, 2002. Marking sense of students' answers to multiple-choice questions. The Physics Teacher 40: 174–180.

Espinosa, M.P. and Gardeazabal, J., 2010. "Optimal correction for guessing in multiple-choice tests." *Journal of Mathematical Psychology* 54, 5: 415-425.

Al-Faris, E.A. et al, 2010. A practical discussion to avoid common pitfalls when constructing multiple choice questions items. *Journal of Family & Community Medicine* 17: 96–102. doi:10.4103/1319-1683.71992

Fellenz, M.R., 2004. Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment & Evaluation in Higher Education* 29:6: 703-719.

George, S., 2003. Extended matching items (EMIs): solving the conundrum. *Psychiatric Bulletin* 27,6: 230-232.

Grunert, M.L. et al, 2013. Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit. *Journal of Chemical Education* 90 (10), 1310-1315 DOI: 10.1021/ed400247d

Haladyna, T.M., 2004. Developing and Validating Multiple-Choice Test Items. New York: Routledge.

Harper, R., 2003. Multiple-choice Questions — A Reprieve. *BioScience Education* 2, 1: 1-6.

Hemmati, F. and Ghaderi, E., 2014. The Effect of Four Formats of Multiple-choice Questions on the Listening Comprehension of EFL Learners. Proceedings of the International Conference on Current Trends in ELT. *Procedia - Social and Behavioural Sciences* 98: 637–644. doi:10.1016/j.sbspro.2014.03.462

Hinchliffe, J., 2014. CQ University scraps multiple choice exams in an Australian first. ABC News, 23 September 2014. Accessed: 28 July 2017. http://www.abc.net.au/news/2014-09-23/cqu-scraps-multiple-choice-exams-in-an-australian-first/5763226.

Ibbett, N.L. and Wheldon, B.J., 2016. The Incidence of Clueing in Multiple Choice Testbank Questions in Accounting: Some Evidence from Australia. *The E - Journal of Business Education & Scholarship of Teaching*. 10, 1: 20-35.

Johnson, J., 2014. Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Rules. Waltham, MA: Morgan Kaufmann.

Kastner, M. and Stangl, B., 2011. Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter? *Procedia - Social and Behavioral Sciences* 12: 263-273.

Lesage, E. et al, 2013. Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation* 39, 3: 188-193.

Kuechler, W.L. and Simkin, M.G., 2010. Why Is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test. *Decision Sciences Journal of Innovative Education* 8: 55–73. doi:10.1111/j.1540-4609.2009.00243.

Masters, J.C. et al, 2001. Assessment of multiple-choice questions in selected test banks accompanying test books used in nursing education. *Journal of Nursing Education* 40: 25–32.

Matters, G. and Burnett, P.C., 1999. Multiple-Choice versus Short-Response Items: Differences in Omit Behaviour. Australian Journal of Education 43: 117–128. doi:10.1177/000494419904300202.

McClusky, H.Y., 1934. The negative suggestion effect of the false statement in the true-false test. *The Journal of Experimental Education* Educational Measurements 2, 3: 269-273

McDermott, K.B., 2006. Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition* 34, 2: 261-267.

McElvaney, E.J., 2010. "Using Constructed-Response and Multiple-Choice Questions in Undergraduate Examinations." *World Journal of Management*. 2: 98-106.

McElvaney, E.J. et al, 2012. Assessment method difference : comparisons between international and domestic students within a first year undergraduate management course. *International Review of Business Research Papers* 8: 205–214.

Moore, H., 2014. Comment on "Does the student a) know the answer, or are they b) guessing?" [The Conversation]. Available at: https://theconversation.com/does-the-student-a-know-the-answer-or-are-they-b-guessing-31893 [accessed April 19, 2018].

Morrison, S. and Free, K., 2001. Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education* 40: 17-24.

Nicol, D., 2007. E-assessment by design: using multiple- choice tests to good effect. *Journal of Further and Higher Education* 31: 53-64.

Prieto, G., and Delgado, A. R., 1999. The effect of instructions on multiple-choice test scores. European Journal of Psychological Assessment, 15(2), 143–150.

Remmers, H.H. and Remmers, E.H., 1926. The negative suggestion effect of true-false examination questions. *Journal of Educational Psychology* 17, 52-56.

Roediger III, H.L. and Marsh, E.J., 2005. The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology* 31, 5: 1155–1159

Roper, M., 2008, 2014. Scientists calculate odd ways to die. *Daily Mirror*, 30 May 2008, updated 3 April 2014. Accessed 8 August 2016. http://www.mirror.co.uk/news/weird-news/scientists-calculate-odd-ways-die-282884.

RWAP Services. n.d. The Chances Of Winning The UK National Lottery. Accessed 8 August 2016. http://playlotto.org.uk/lottery/uklottery_odds.html.

Simkin, M.G. and Kuechler, W.L., 2005. Multiple-Choice Tests and Student Understanding: What Is the Connection? *Decision Sciences - Journal of Innovative Education* 3: 73–98.

Srivastava, A. et al, 2004. Why MCQ. *Indian Journal of Surgery* 66:246-8.

Tarrant, M. et al, 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice* 6: 354–363.

Traub, R.E. and Fisher, C.W., 1977. On the Equivalence of Constructed- Response and Multiple-Choice Tests. *Applied Psychological Measurement* 1(3):355-369 DOI: 10.1177/014662167700100304

Wainer, H. and Thissen, D. 1993. Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education* 6: 103–118.