To Choose or not to Choose Multiple-Choice

Abstract

In this study we analyze the outcome of introducing Multiple-Choice (MC) questions to assess engineering students' ability to apply basic concepts in previously unfamiliar problems (in the field of semiconductor devices). Using Item Response Theory (IRT) it is observed that MC questions of the design we have employed can work very well, while at the same time presenting a risk of resulting in test items that are worse than any of the Constructed Response (CR) questions they are compared to. We argue that MC test items intended for assessment of engineering skills have a higher need of proper validation before use than CR problems.

Introduction

Using Multiple-Choice (MC) questions as part of the examination of students' abilities can save a substantial amount of time when correcting their answers (Lukhele, 1994), enabling *e. g.* increased time for other forms of interaction between teachers and students. It has been shown, however, that the quality of the measure of a particular ability when using MC questions depends not only on how well the MC items are constructed and graded (Lesage, 2013), but also on the nature of the student ability that is to be measured (Ward, 1980)(Ward, 1987).

The literature on MC testing includes both general recommendations for MC test item design (Rodriguez, 2013) as well as suggestions applied in engineering contexts (Triantis, 2013) (Farthing, 1998). We have made our own design and implementation of MC test items in examinations in a compulsory course for third year electrical engineering students. The students have access to some 200 MC questions of various formats for training during the course, but in the final written examination we implemented a particular MC design to test the fulfillment of the learning outcome of being able to *identify the applicability of basic subject concepts for making reasonable inferences in simple but unfamiliar problems*, a skill which we otherwise typically assess by evaluating some form of Constructed Response (CR) from the student.

Scoring CR answers requires quite some effort. The CR question typically asks for an explanation, motivation or argument, and there is seldom a single correct answer. The students' writing needs to be interpreted and, for a fair scoring, the judgements of all the individual and unique answers need to be calibrated. We deal with this by formulating a concise grading rubric after reading through all CR answers handed in by the students, making sure that it is applicable in a fair way to every solution. The time gained by using automatically scored MC questions is thus substantial in our case, and we would be able to make good use of a digital examination format. What we do need to consider is that the quality of assessment provided by the MC format is at least comparable to using CR, and that the time gained in scoring is not all lost in added time for test item construction. These two aspects are addressed in this paper and our intention is that the results will support fellow

engineering educators to make better choices when considering using MC tests in their courses.

Background and design of study

In order to make inferences regarding the consequences of modifying the examinations to include the MC test items we do the following:

1. use Item Response Theory (IRT) (Sijtsma, 2006) to compare the outcome for MC test parts with the outcome for similar CR items in two recent examination events, comprising near 100 students
2. estimate the average effort (in time) to construct and grade either MC or CR test items

In IRT it is assumed that all items (questions) in a particular test are giving a measure of one and the same trait. The outcome of the entire test can be used to designate a trait-value for each test taker. It is then possible to derive an Item Response Function (IRF) for each test item, which gives the average score (or likelihood for a correct answer) on the test item as a function of the value of the trait. In a test that is to assess the ability of each test taker by adding up the individual's scores of all test items, we would like each IRF to be monotonically increasing with ability and also to be showing a high degree of discrimination between low and high ability. Knowing the IRF of all test items, the interpretation of the outcome of the test can be improved *e. g.* by giving more weight to good test items and by disregarding bad ones. Figure 1 shows three examples of qualitatively different IRFs. Although the actual average scores calculated from a finite sample only constitute *estimates* of the IRF score values for each value of ability, we will refer to the set of data points for a question appearing in a graph as that in Figure 1 as an IRF.

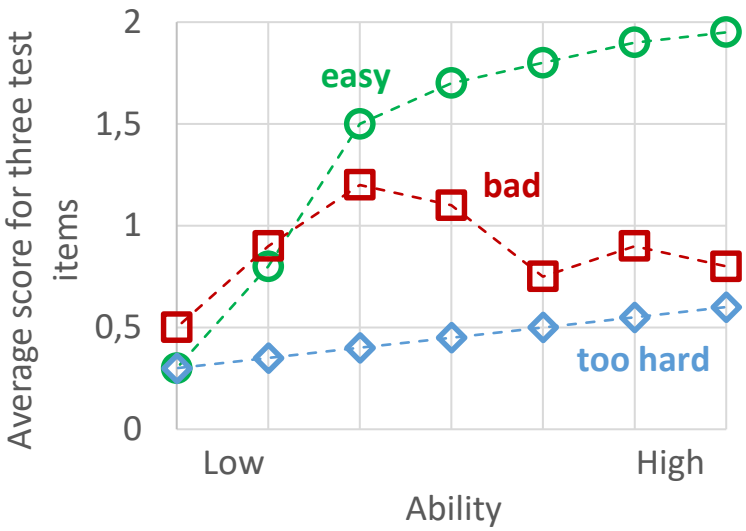

Figure 1. Three illustrative Item Response Functions: the green circles represent an easy question (even rather low ability gives a high average score), the question with blue diamonds is too hard (giving poor discrimination for ability), and the bad red squares is for a question where higher ability does not result in a better score.

The course context in which this study is conducted is the following: the total credits of the course are 7,5 of which 3 are examined in an end-of-the-course written exam. The remaining credits are examined in project work and assignments during the quarter the course is running. The exam is divided into two parts – problems 3-5 rely on the students' access to and proficiency in using reference literature (the course text book) and often deal with data presented in graphs taken from recent research publications. Each of these problems can give up to 4 points. Problems 1-2 need to be solved with only paper and pencil. Problem 2 can give 2 points and is related to a particular sub-topic in the course which to a large extent concerns memorizing facts. Problem 1 contains five parts (a-e) where each part is a separate question which can give 2 points. The student must choose which four of these five parts she wants to address, making 8 the maximum score for problem 1. In order to pass, the student must have a total score of at least 9 on the exam, and at least 5 points must come from problem 1.

All the questions in problem 1 are designed so that a student with passable ability should obtain a full score for all of them. Since there is a risk that a student actually has an adequate ability, except for handling one particular question on an exam, the student gets to de-select one of the five parts. There is furthermore the risk of a student misunderstanding a question in such a way that she cannot demonstrate her ability truthfully in its solution, wherefore it is only required that three of the four selected problems are satisfactorily handled – which would correspond to 6 points in total. Allowing for significant uncertainty in the discrimination between what is required for a full score (2 points) and the (only) partial score (1 point, since we count only whole points) for one question that results in the lower score for any one student, the final score requirement to pass ends up at 5 points. In reality, very few students of those who obtain 5 points in problem 1 fail to meet the requirement of 9 points in total (<10%). The way we have set up problem 1 makes it less sensitive to a single bad question – the students have a chance to de-select it, and they have a margin to fail it and still pass.

The format of the MC test items we introduced contains a brief introduction (which might include a graph) and then gives four statements about the described situation. Below is an example of the format:

Select which of the two alternatives below that are most obviously reasonable when constructing exam problems (deduction is done for improper choices and you can select three or one alternative for 1 point. The lowest score is 0):

1. *the problem should be confusing for the students*
2. *the problem should test if the student is smart*
3. *the problem should address one of the learning outcomes of the course*
4. *the problem should be possible to answer incorrectly*

The four alternatives are here not independent proposals of answers to a question, but rather individual statements, which either could be "better" or "worse" than the other alternative statements; so the MC format is more appropriately described as four linked binary questions (of "true" or "false" character) rather than one question with four answer alternatives. It constitutes a special case of what has been referred to as Multiple True False (MTF) questions, which have been reported to show higher reliability than ordinary MC questions (Haladyna, 2002). A good design requires the (challenging) formulation of four statements that are neither trivially wrong nor trivially correct and at the same time sufficiently well described to allow passing solid judgement on. The slightly elaborate format is a consequence of the intention of assessing the ability to *identify applicability of concepts* rather than e g

being able to correctly perform a calculation or to know the definition of a concept. In the example above alternatives 3 and 4 are (supposed to be) the most obviously reasonable ones. The next example is constructed to demonstrate the format in an engineering context and with statements of poor quality:

Select which of the two alternatives below that are the most obviously reasonable consequences of applying a constant downwards directed force, F, to the free end of a horizontally mounted cantilever (deduction is done for improper choices and you can select three or one alternative for 1 point. The lowest score is 0):

1. *the cantilever will break*
2. *the cantilever will tend to move*
3. *the free end of the cantilever will be displaced upwards*
4. *the acceleration is proportional to the applied force*

The statements have the following problems: 1 – this *could* be reasonable if the force is strong enough, so the formulation is insufficient for passing a solid judgement. 2 – this vague formulation is trivially correct. 3 – is trivially false. 4 – this is a formulation derived from Newton's second law, so it is in many cases a verifiable (true) statement, but it is not formulated as a consequence of the conditions described; it is difficult to interpret how to relate to this statement in the given context. If we instead had used a CR approach and just asked for a (brief) description of the most reasonable consequences, we would have saved considerable design effort and avoided ending up with a question which due to its low-quality alternatives would give very poor information on the actual ability of the students. Many students with poor ability would settle with selecting alternative 2 (scoring 1 point), whereas a higher performing student selects alternatives 2 and either 1 or 4, one of which would result in a net score of 0 points.

Dealing with blind or informed guessing and partial knowledge when scoring MC tests is an issue without simple solutions (Lesage, 2013). One recent suggestion is to use the outcome of (partial) elimination of the least probably alternatives to form a measure of an ability (Wu, 2019), which is argued to provide a better measure than just considering the selection of correct alternatives. In our design we have addressed these issues through giving the students the opportunity to "play safe" and receive a half score (1 point) if they identify either one of the most reasonable or one of the most improper alternatives (and select one or three alternatives). For blind guessing the best option is to just select one random alternative, which would result in the average score of 0,5 points.

Measuring the time spent on constructing and correcting questions with a stopwatch while "in the act" of doing it has not been carried out in the frames of this study. We will however argue that a decent first order approximation for the difference in time required between CR and MC questions in our case can be obtained by reasoning about the structural differences in formats.

Results and discussion

Our IRT analysis indicates that the MC test items are more sensitive to test item design – a good MC test item can display a near ideal IRF, whereas a badly designed one is worse than the worst CR item. Figure 2 shows the IRF for two MC items and IRFs for the two most extreme CRs in the recent examinations (selected from a total of seven questions). A good

IRF increases monotonically from low score to high score – it is crucial that a higher total score should imply a higher average score for the individual test items; the MC item given in red squares deviates significantly from this behavior, meaning that it is a badly functioning test item. For the two CR items one appears to be too easy, resulting in rather poor discrimination (the blue diamonds) and the other (purple triangles) is a bit too hard. The green circle MC item looks very nice, with high discrimination and appropriate level of difficulty.

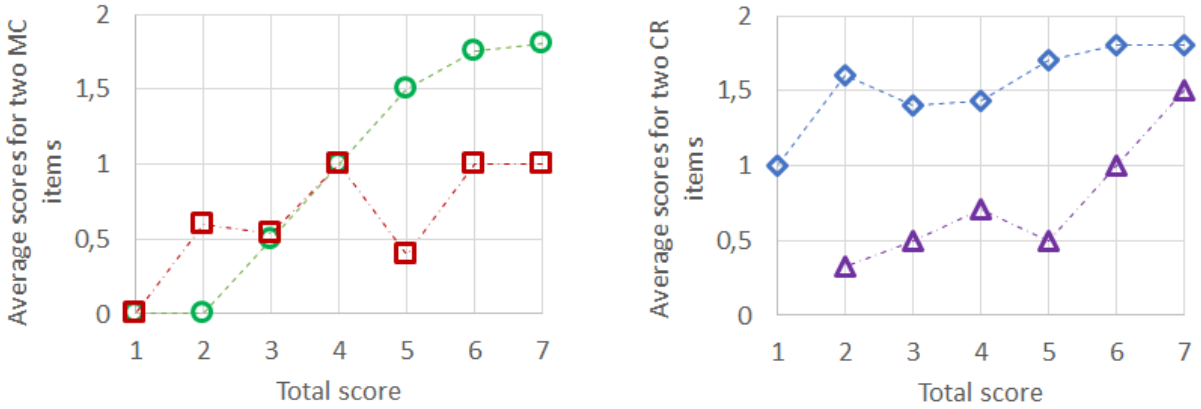

Figure 2. IRFs for two MC items (green circles and red squares) and the two most extreme CR items out of a total of seven (blue diamonds and purple triangles).

Since the number of students with the different total scores varies, the reliability of the item score averages to estimate the IRF also varies. In this particular data set there were no students with a full score (8 points), but had there been, their average result for any item would have had to be 2 (just as 0 total points means a 0 average for all items). In the graphs in Figure 2 the red square curve is the one with the biggest skew in uncertainty between values; the item averages for a total score of 3 and 5 are the most reliable ones (more than 13 samples for each). In view of our grading scheme it is particularly dissatisfying that the ability measure deemed as sufficient (total score of 5 points) results in an average failure on the question and a *lower* value than that of those ending up with a non-passable ability (total of 3 points). One possible explanation to such an outcome could be that less able students choose to guess at only the one most likely alternative (which seems to be too easy to spot), whereas the more able students select one more alternative, where the (confusing) difficulty to pick the correct one makes many of them choose incorrectly.

In a study by Peuker *et al*. (Peuker, 2013) CR and MC versions of the same test items, concerning simple mechanics related calculations, were evaluated in an introductory engineering course with 75 students. Their findings are very similar to ours, *i. e.* that MC questions can overall function equally well as CR, but there is a risk of ending up with significant discrepancy between MC and CR results for individual questions.

The negative consequences of a single poorly designed MC test item are more severe than for problems requesting a CR. The limited outcomes of the MC test items imply that very limited information is obtained from the student, and the quality of the information received is determined entirely by the design of the question and answer options; as an examiner you can discriminate whether the student "understands" *X* or not - if the test item is well designed. In CR even the worst of questions can actually generate useful and scorable information from the student (show that the student knows *Y* or *Z*).

In the case where you can apply a very large number of test items, the quality of each individual item is less critical. Our starting point was an existing examination format that we believe functions well to assess the appropriate abilities. In the process of replacing the CR format for a more easily scored MC format, we need to maintain fidelity in assessing the same thing – which prompted the particular MC design we adopted. As we will argue below, the time required for each item construction increases when shifting from CR to MC, so further increasing the construction time for the whole exam by introducing more questions does not appear immediately attractive. In this context, also the time cost for the students in solving the problems must be taken into account, which is something we have not included in this study.

With regards to the effect on the time spent by the examiner on exam construction and grading when shifting from CR to MC it is important to first state that in the opinion of the author, the construction of problems for an exam is a process requiring significant creativity. Our standard is to never re-use or simply re-formulate problems that have been used before. Constructing a new examination that will properly assess the same competences as all previous exams, but with completely new questions, takes time.

Assuming that the required time has been spent to arrive at a satisfactory CR question – how much extra time would we need to turn this into a MC item? In our format, the conversion would in principle imply adding four explicitly formulated options that reflect the content of the problem in a meaningful way. We now argue that the formulation of the CR question in the first place requires at least making sure that: 1) there is a viable and reasonable approach to the question (providing one alternative for the MC version) and 2) the question is challenging enough to inherently provide a risk of making (at least) one known mistake. This is to say that two candidates of the four alternatives for the MC question should present themselves rather immediately. As has been shown in the previous results, the creation of viable true/false statements for a MC question is not straightforward; each new alternative considered needs to be balanced with the existing ones. The overall estimate from our experience is that the creative act of defining a CR type questions takes us half-way to our MC version (it provides two alternative statements) – and the remaining process to find two more is of similar creative magnitude.

Whatever format of problems on the exam, it should be subjected to some form of quality check. A peer review process has been demonstrated to significantly improve the quality of MC examinations in medicine (Malau-Aduli, 2012), and considering the results of our study, it appears that MC questions are likely to have more to gain in such a review process than CR type questions. Ideally, each MC item should be tested towards a representative audience to have a first indication of its IRF. In view of the high demand of resources required to implement such a scheme, a simple collegial discussion around the questions would go a long way to avoid the worst mistakes from degrading the quality of the exam.

The balance in the equation of cost of resources directly depends on the time saved in grading when employing the MC format. We end up with an estimate that the time for construction is doubled (in our case going from some 4-8 to some 8-16 hours for constructing problem 1 on the exam), and the time saved in grading 50-60 students is of the same order; the grading of problem 1 comprises reading and interpreting all student answers (estimated at requiring 5 minutes per student) *and* formulating a valid grading rubric, which requires sufficiently careful analysis of all the constructed answers and of their differences. This analysis does not scale directly with the number of students, and we estimate it to take approximately 1 hour for a typical exam of more than 20 students. In the cases where we have exams with significantly

fewer participants than 50, it is really doubtful whether we save time by shifting over from CR to MC.

Conclusion

Replacing CR questions with equivalent MC questions can result in maintained quality at a reduced cost. In order to achieve this outcome, it is necessary to ensure that the time saved in grading is greater than the extra time required to construct a MC question. In our case we need more than 50 students in the exam to have a grading time gain that compensates for the near doubling of time in problem construction. Furthermore, seeing that there appears to be an increased risk of formulating problematic questions, we recommend that the quality of MC test items intended to assess engineering skills and abilities that are at least not of the most basic kind (recall) should be properly validated before using them in an examination. If this kind of validation is not already conducted for CR questions, this implies a further added time cost.

References

Farthing, D. W., Jones, D. M. and McPhee, D, "Permutational Multiple-Choice Questions: An Objective and Efficient Alternative to Essay-Type Examination Questions", presented at ITiCSE'98, *Innovation and Technology in Computer Science Education-1998*, Dublin, Ireland.

Haladyna, T. M., Downing, S. M. and Rodriguez, M. C., "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment", *Applied Measurement in Education,* v. 15 (2002), pp. 309-334.

Lesage, E., Valcke, M. and Sabbe, E., "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*, v. 39 (2013), pp. 188-193.

Lukhele, R., Thissen, D. and Wainer, H., "On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests", *Journal of Educational Measurement*, v. 31 (1994), pp. 234–250. doi:10.1111/j.1745-3984.1994.tb00445.x

Malau-Aduli, B. S., Zimitat, C., "Peer review improves the quality of MCQ examinations", *Assessment & Evaluation in Higher Education*, v. 37 (2012), pp. 919-931. doi: 10.1080/02602938.2011.586991

Peuker, J., McFerran Brock J. and Peuker S., "Effect of Multiple Choice Testing on Student Performance in an Introductory Engineering Course", *2013 ASEE Annual Conference & Exposition*, Atlanta, Georgia. https://peer.asee.org/1947

Rodriguez, M. C., "Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research", *Educational Measurement: Issues and Practice*, v. 24 (2005), pp. 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Sijtsma, K. and Junker B. W., "Item Response Theory: Past Performance, Present Developments, and Future Expectations.", *Behaviormetrika*, v. 33 (2006), pp. 75-102.

Triantis, D. et al., "Comparing Multiple-Choice and Constructed Response Questions Applied to Engineering Courses", *Communications in Computer and Information Science*, v. 510 (2013), pp. 130-140.

Ward, W., Frederiksen, N., and Carlson, S., "Construct Validity of Free-Response and Machine-Scorable Forms of a Test", *Journal of Educational Measurement*, v. 17 (1980), pp. 11-29.

Ward, W. C., Dupree, D. and Carlson, S. B., "A Comparison of Free-Response and Multiple-Choice Questions in the Assessment of Reading Comprehension", *ETS Research Report Series*, 1987: i–26. doi:10.1002/j.2330-8516.1987.tb00224.x

Wu, Q., De Laet, T. and Janssen R., "Modeling Partial Knowledge on Multiple-Choice Items Using Elimination Testing", *Journal of Educational Measurement*, v. 56 (2019), pp. 391-414.