Practical Assessment, Research, and Evaluation

Volume 23 *Volume 23, 2018* Article 18

2018

Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions

Lucia Otoyo

Martin Bush

Recommended Citation

Otoyo, Lucia and Bush, Martin (2018) "Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions," *Practical Assessment, Research, and Evaluation*: Vol. 23 , Article 18.
DOI: https://doi.org/10.7275/hq8a-f262
Available at: https://scholarworks.umass.edu/pare/vol23/iss1/18

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 18, December 2018 ISSN 1531-7714

Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions

Lucia Otoyo and Martin Bush
London South Bank University

This article presents the results of an empirical study of "subset selection" tests, which are a generalisation of traditional multiple-choice tests in which test takers are able to express partial knowledge. Similar previous studies have mostly been supportive of subset selection, but the deduction of marks for incorrect responses has been a cause for concern. For the present study, a novel marking scheme based on Akeroyd's "dual response system" was used instead. In Akeroyd's system, which assumes that every question has four answer options, test takers are able to split their single 100% bet on one answer option into two 50% bets by selecting two options, or into four 25% bets by selecting no options. To achieve full subset selection, this idea was extended so that test takers could also split their 100% bet equally between three options. The results indicate increased test reliability (in the sense of measurement consistency), and also increased satisfaction on the part of the test takers. Furthermore, since the novel marking scheme does not in principle lead to either inflated or deflated marks, this makes it easy for educators who currently use traditional multiple-choice tests to switch to using subset selection tests.

For high quality assessments, the key issues are test reliability and validity/fairness (Caines *et al.*, 2014). Traditional multiple-choice tests are so familiar that they are generally taken for granted, but their conspicuous dependency on chance makes them a relatively unreliable assessment method. The overall impact of chance can be reduced by including more questions within a test, but this involves more work for both the test creator and the test takers. One alternative is to require the test takers to take a given test twice in quick succession, and to then calculate their average scores; this would also involve more work for the test takers, but no additional work for the test creator if the marking is automated. This would not be a very satisfactory solution from the point of view of the test takers, however.

This article discusses the merits of the subset selection test format, which is a generalisation of the traditional multiple-choice test format in which the impact of chance is reduced by enabling test takers to select multiple answer options whenever they are unsure which option is correct for a given question. It describes an empirical study involving a novel marking scheme that does not involve any deduction of marks, in contrast to the conventional marking scheme for subset selection tests in which a fixed penalty is incurred for each distracter (i.e. wrong answer) selected.

The content of this article should be of interest to all designers of multiple-choice tests, especially in the context of high-stakes summative assessment, where the reliability (in the sense of measurement consistency (Tavakol and Dennick, 2011)) of test scores and the stress induced by the testing process are both major concerns.

Background

Traditional multiple-choice tests

Traditional multiple-choice tests originated over 100 years ago (Suen and French 2003). Today they are

very widely used at all levels of education. The most popular marking scheme for such tests is often referred to as *number right scoring*, since one mark is awarded for each correct answer.

They are a rather crude method of assessment, however. To illustrate this, consider the chess question shown in Figure 1.

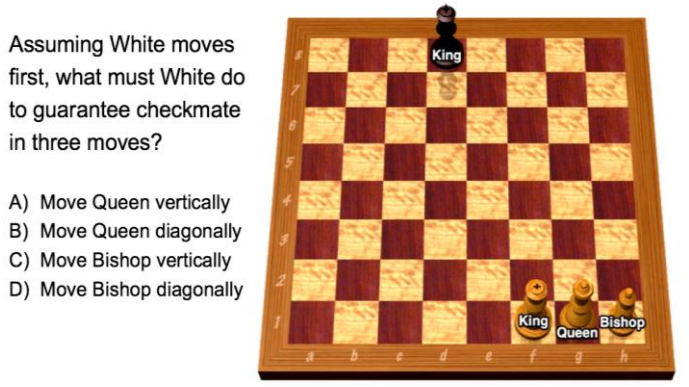

Figure 1. A multiple-choice question relating to the game of chess

Suppose that a group of test takers are asked to answer this question. It is clearly not reasonable to assume that all those who select the right answer (A) are good chess players. There may be some who select the right answer through guesswork.

Traditional multiple-choice tests with "negative marking"

Negative marking – i.e. deducting marks for incorrect responses – acts as a counterbalance to score inflation due to pure guesswork. That is why it is often referred to as "correction for guessing" (Diamond and Evans 1973). It is quite widely used (Zapechelnyuk 2015), but it is controversial because it may inhibit some test takers from answering questions correctly when they are not fully confident (Betts *et al.*, 2009).

The usual formula is as follows:

$$S = R - W/(n\text{-}1)$$

where S is the corrected test score, R is the number of right answers selected, W is the number of wrong answers selected, and n is the number of answer options per question (which must be consistent throughout the test).

Use of this formula to counterbalance pure guesswork does nothing *in itself* to improve the reliability of multiple-choice tests. In the words of Wood (1991, p41), "That undeserved reward can be obtained through blind guessing is a permanent blemish on multiple choice, one that is not removed by so-called guessing corrections." However, if the test takers are made to understand that they are just as likely to lose marks as to gain marks when they engage in pure guesswork, then this should inhibit pure guesswork to some extent.

Consider the chess question shown in Figure 1 once again, and in particular how this question might be answered by test takers who have no knowledge of chess whatsoever. When negative marking is used, one would expect fewer of those test takers to attempt it. Indeed, some researchers have found that risk-averse test takers do often choose to omit questions rather than engage in pure guesswork (Avila and Torrubia 2004; Betts et al., 2009). The key point here is that less guesswork implies more reliable test scores (Karandikar 2010).

On the other hand, there may be some *novice* chess players amongst the test takers who know that C is an illegal move although they are unable to determine, within the time available, which of A, B or D is the correct answer. Negative marking is designed to discourage *pure* guesswork, but since these novice chess players are able to eliminate option C they are in a position to make an *educated* guess – therefore they have more to gain than to lose (on average) by attempting to answer the question.

Conventional subset selection tests

Whilst pure guesswork may be the most obvious blemish on multiple-choice tests, *educated* guesswork also clearly harms test reliability. Subset selection tests – which date back at least 65 years (Dressel and Schmid 1953) – address this problem by enabling test takers to select any subset of the answer options for each question. It is often said that subset selection tests enable test takers to express *partial knowledge* (Ben-Simon *et al.*, 1997); in other words, they can indicate explicitly that they are able to successfully eliminate at least one distracter.

Subset selection tests are designed to yield more reliable test scores than traditional multiple-choice tests because the test takers are no longer required to choose arbitrarily (i.e. to make a guess) between alternative answers which they favour equally.

The conventional marking scheme for subset selection tests is based on that of traditional tests with

negative marking (Jaradat and Sawaged 1986), and is as follows, where n is the number of answer options per question:

- for each correct answer selected => award *n*-1 marks
- for each incorrect answer selected => deduct 1 mark

For a test with four options per question – commonly referred to as a *4-choice test* – the marking scheme is therefore as follows:

- for each correct answer selected => award 3 marks
- for each incorrect answer selected => deduct 1 mark

Conventional subset selection tests reward partial knowledge more generously than traditional tests do. For example, an intermediate chess player who can eliminate distracters C and D in the question shown in Figure 1, but who cannot decide between A and B (in the time available), is able to select both A and B – with the subsequent reward being a mark of 2 out of 3 – i.e. 0.67. By contrast, in a traditional test their mark would be 0.5 *on average* because they would be forced to choose arbitrarily between A and B.

In other words, 4-choice conventional subset selection tests are likely to benefit test takers who can often eliminate at least two distracters per question within a given test (Jennings and Bush 2006). On the other hand, test takers who are often unable to identify any distracters whatsoever may end up with a lower score by comparison with what it would be in a traditional multiple-choice test. In the extreme, a test taker who can do no better than resort to pure guesswork for every question would have an expected (mean) score of 0% in a conventional subset selection test, versus an expected (mean) score of 25% in a traditional multiple-choice test.

This can make it difficult for educators who would like to switch from using traditional multiple-choice tests to conventional subset selection tests, especially in situations where it is important to retain consistency in terms of marks and grades with respect to previous cohorts of students.

The possibility of losing marks due to negative marking also means that conventional subset selection tests are likely to be stressful for some test takers (Bush 2001). The impact of the negative marking can appear very punitive, since the mark per question in a 4-choice test ranges from +3 (when only the correct answer is selected) down to –3 (when all three distracters are selected).

Subset selection without mark deductions

The two concerns identified above in relation to conventional subset selection tests – i.e. test taker stress and the inconsistency of scores with respect to traditional tests – are both alleviated by using an alternative marking scheme that does not involve any deduction of marks, as follows (for a 4-choice test):

- correct answer only selected => 1 mark
- correct answer plus one distracter => 0.5
- correct answer plus two distracters => 0.33
- all answers (or no answers) selected => 0.25
- any other response => 0.

In the absence of guesswork, the second and third of the five bullets points above both correspond to expression of partial knowledge, whilst the fifth bullet point corresponds to the expression of *misinformation* (Bradbard *et al.*, 2004). The marks awarded have been chosen to coincide exactly with the expected (mean) mark for the same question in a traditional test:

- choosing the correct answer only => 1 mark
- choosing arbitrarily between 2 answers => 0.5 on average
- choosing arbitrarily between 3 answers => 0.33 on average
- choosing arbitrarily between all 4 answers => 0.25 on average
- choosing any answer but the right one due to misinformation => 0.

Hence this new marking scheme should in principle yield test scores that are consistent (i.e. neither inflated nor deflated) with respect to those resulting from a traditional test. This implies that individual test takers are neither more nor less likely to pass, irrespective of the pass mark. Assuming ideally rational test taker behaviour, it ensures an equivalent statistical profile of

test scores with respect to previous cohorts of test takers who were assessed using a traditional 4-choice multiple-choice test.

This new marking scheme was first described, to the best of the authors' knowledge, in a 2014 TEDx talk (Bush, 2014). It is based on Akeroyd's *dual response system*, which only allows test takers to choose one or two (or no) answers per question (Akeroyd 1982).

A simple way of explaining the new marking scheme to test takers is to point out that by selecting more than one option they are effectively splitting their 100% bet on one answer option into two 50% bets, or three 33% bets, or four 25% bets. Looking at the new marking scheme in this way makes it clear that it is in fact a restricted form of what has previously been called *probabilistic* tests (Rippey 1968) and *multiple-evaluation* tests (Holmes 2002), in which the test takers are required to associate a percentage 'personal probability' to every answer option for each question.

The following generalised description for n-choice tests utilising this new marking scheme has been proposed by Zapechelnyuk (2015):

- any number, k, of the n answers may be selected
- selection of no answers is regarded as selection of all answers
- a mark of $1/k$ is awarded if one of the selected answers is correct, otherwise zero is awarded.

The empirical study

Design

A cohort of 75 first year undergraduate students took three summative assessments, each one covering a third of the syllabus of their introductory one-semester course on computer architecture. The students were all undertaking a BSc in Information Technology or a closely related discipline. The one-semester course covered topics such as data representation using binary numbers, CPU operation, assembly language programming, volatile memory and non-volatile data storage technologies, data transmission, computer networking and operating systems.

These 75 students were divided into five groups for their tutorial classes – to make the class sizes manageable – and also, therefore, for the assessments. When taking the assessments they were required to answer every test question twice before moving on to the next question; once according to the rules of a traditional multiple-choice test (where only one answer can be selected per question) and once according to the rules of a subset selection test (where multiple answers can be selected per question). The marking schemes – i.e. number right scoring for the traditional tests, and the novel marking scheme described above for the subset selection tests – were carefully explained, and the students were given practice tests prior to the first assessment to familiarise them with subset selection and the dual test format. The three assessments were then administered at roughly equal intervals over a period of one academic semester.

With 75 students each supposed to take three assessments, there could potentially have been a total of 225 assessments included in the study. However, the students didn't all turn up for every test, nor did they give their consent on every occasion for their test responses to be included anonymously in the study.

Each assessment involved 20 multiple-choice questions – the same 20 questions were presented, in the same order, to every student – and each question had to be answered twice as described above. The questions were designed by a senior academic with many years of experience of creating multiple-choice questions in the area of computer architecture. Most of the questions used were revised versions of questions that had been used in previous years. All of the questions were peer reviewed by the teaching team.

The students were told in advance that whichever test type yielded the highest average mean score for the whole cohort for each assessment would be the one that counted for all of the students for that assessment. This ensured that the students would try their best to optimise their marks for both test types.

Finally, the students were invited to complete an anonymous online questionnaire after the last assessment (but before seeing their test scores); its aim was to find out what the students thought about each of the test types.

Ethical considerations

Since this empirical study gathered data from summative assessments that were compulsory for all students, the ethical issues associated with this had to be very carefully evaluated. Measures were put into place to safeguard the students' interests. The main concern was that the performance of some students might be

adversely affected by the dual test requirement. It was important to be able to determine whether underperformance in any particular case was due to the testing procedure or the student's lack of preparation.

Adverse effects were mitigated as follows:

1. All students were fully informed about the research two weeks prior to the study being conducted, and before each of the three assessments. They were provided with detailed information verbally and in writing, and they were given several opportunities to discuss any concerns.
2. Although the assessments were compulsory, the students were asked to give their consent on each occasion for their test responses to be included anonymously in the study. They were able to opt out. Out of the 216 assessments taken, 16 had to be excluded from the analysis for this reason, leaving exactly 200 that could be analysed.
3. A pilot study was conducted two weeks prior to the first assessment, which provided sufficient time to make adjustments to the wording of the instructions etc.
4. The students were given 10 minutes extra to complete the dual tests versus the normal time limit (40 minutes) for similar 20-question tests given in previous years.
5. A mock assessment was given to all students a week before the first real assessment.
6. Counting the scores from whichever test type yielded the highest average score for the cohort benefitted the students overall, but it did not necessarily benefit every student. It was decided that in cases where a student failed an assessment which they would have passed had their score for the other test type been counted, their score would be raised to a bare pass for that assessment. (As the study progressed, this only had to be done once.)

With these mitigations in place, and bearing in mind that the first year marks did not affect the final degree classification, it was concluded that the study would not significantly disadvantage any student.

Results

Table 1 presents test pair statistics for the three dual tests: the mean score (out of 20) and the standard deviation for each of the three pairs of tests (*Traditional* and *Subset Selection*), the test score pair correlations, and the sample sizes.

Table 1. Test pair statistics for the three dual tests

	Mean test scores		Standard deviations		Test score pair correlations	Sample size
Dual test	Trad. test	SS test	Trad. test	SS test		
1	10.85	10.50	2.95	2.94	0.913	65
2	10.62	10.26	3.44	3.13	0.919	60
3	10.52	10.52	3.88	3.72	0.915	75

As the data in Table 1 shows, the students performed slightly better on the whole when taking each traditional test versus the equivalent subset selection test. Out of the 200 assessments included in the study, a higher score was achieved in the traditional test on 101 occasions versus a higher score in the subset selection test on only 52 occasions.

The time that was allocated for completion of the tests turned out to be more than enough for the vast majority of the students. Some students with certain recognised disabilities (including dyslexia) are routinely given 25% extra time for all university assessments, and so there were some students for whom the time limit was 50 minutes rather than 40 minutes. Only a very few students stayed until the end of each test, and it appeared that they had in fact answered all of the questions within their allocated time limit but that they were using whatever remaining time they had to review their answers prior to submission.

These results were no surprise. The marking scheme for the subset selection tests was designed to yield the same mean scores as traditional tests, but slightly higher marks for the traditional tests are to be expected, since risk averse test takers may sometimes 'play safe' by selecting both their first-choice and their second-choice answer options to one or more questions. Whenever their first-choice option was the correct one, they would have got a higher mark in a traditional test.

Reliability

Cronbach's alpha coefficient, commonly used for estimating psychometric test reliability, was calculated

for each of the six tests; the results are shown in Figure 2.

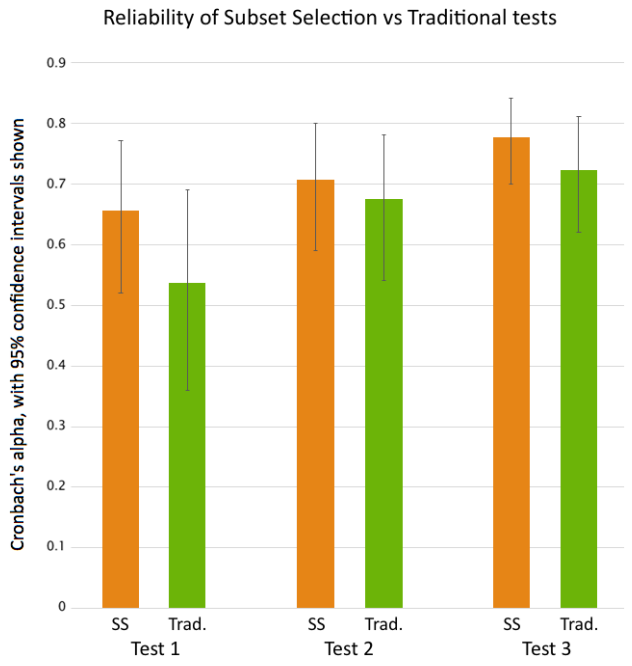

Figure 2. Cronbach's alpha coefficient for the six tests (three dual tests)

Values of alpha (α) can vary from 0 to 1; higher values are more desirable. There are no agreed interpretations of the scales, but Lance at al (2006) provides a guide:

$$\alpha > .9 \rightarrow \textit{Excellent}; \quad \alpha > .7 \rightarrow \textit{Acceptable};$$
$$\alpha > .6 \rightarrow \textit{Questionable}; \quad \alpha < .5 \rightarrow \textit{Unacceptable}$$

As Figure 2 shows, the reliability of each of the subset selection tests appears to be better than the reliability of the corresponding traditional test. However, when the usual 95% confidence intervals are considered (as shown in Figure 2) it becomes clear that there is quite a wide spread due to the relatively small sample sizes. To achieve significantly narrower confidence intervals it would have been necessary to have a much larger cohort of students.

Frary (1989) and Ben-Simon et al (1997) reviewed and compared numerous empirical studies of conventional subset selection tests undertaken in the period 1953–1989; they found that the majority confirmed enhanced reliability with respect to traditional multiple-choice tests. The results from the present study seem consistent with reasonable expectations in this respect.

Which test type did the students prefer?

All 75 students were invited to complete an anonymous online questionnaire after the last assessment; 50 chose to do so. Their feedback, summarised in Figure 3, indicated a clear preference for subset selection tests. Also, 44% of the respondents agreed that they felt less stressed when taking the subset selection tests versus 24% who felt less stressed when taking the traditional tests.

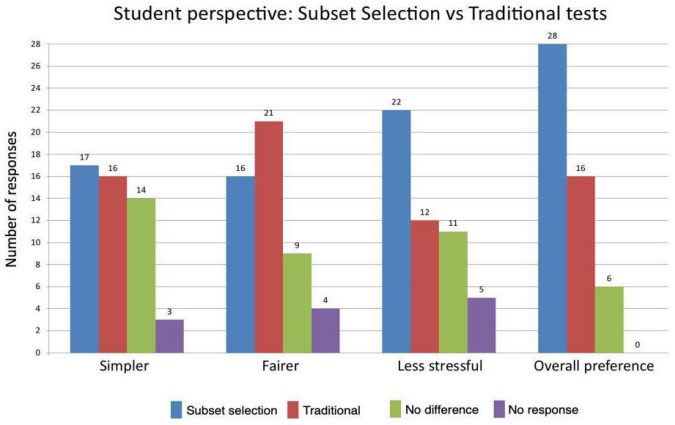

Figure 3. Student feedback

Follow-up survey about attitudes to guesswork

The empirical study described above was followed up in a subsequent year – for another cohort taking the same one-semester course – with a very brief questionnaire that the students were invited to fill in immediately after taking the second of their three tests. This second cohort of 61 students were also assessed using subset selection tests with the novel marking scheme described in section 3 above, but they were not required to take the dual tests described in section 4.1. The test questions were very similar as those used within the dual test taken by the first cohort studied, but not the same.

Due to unusually high absenteeism, only 53 of the 61 students took the second test. 43 of those 53 completed the online questionnaire. The summary of responses to the question "In the test that you have just taken, to what extent did you consciously pick one answer at random between two or more of the answer options when you were unsure which answer was correct?" are shown in Figure 4.

In an effort to keep the questionnaire very simple, and therefore to encourage students to complete it, it was decided not to ask any more detailed questions.

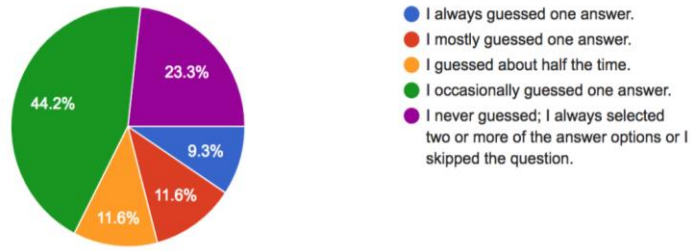

Figure 4. Students' attitudes towards making guesses when taking the novel tests

From the responses shown in Figure 4 it is evident that according to the students' responses they mostly felt inclined not to make guesses. By contrast, if they had been taking traditional multiple-choice tests, as explained in section 2.3, they would have been forced to make a guess whenever they felt unable to identify a preferred option.

Analysis of student responses when taking subset selection tests

The test responses of the 61 students in the second cohort were subsequently analysed in order to see how frequently they were in practice choosing to select none, one, two or three of the four answer options. (Choosing all four options, which students very occasionally did, was regarded as choosing none.) The results are shown in the form of pie charts in Figure 5; they are labelled 1A, 2A and 3A. The test responses for a third independent cohort of 68 students taking the same one-semester course, assessed using the same novel subset selection tests but containing different questions once again, were also analysed; these are labelled 1B, 2B and 3B in in Figure 5.

The pie charts clearly show that when taking the 4-choice subset selection tests:

- the students most frequently selected exactly one answer per question
- the next most frequent response was selection of two answer options
- the next most frequent response was selection of no answer options
- the least frequent response was selection of three answer options

This is in accordance with the authors' experience of using subset selection tests in other years and with other student cohorts.

Eleven of the 53 students in the second cohort who took the second test (i.e. roughly 20%) selected exactly one answer to every question. Since 9.1% of the 43 survey respondents from that group of 53 had said that they "always guessed one answer" (see section 4.6 above), this indicates, roughly speaking, that around half of the 20% consciously engaged in guesswork while the other half did not. Those who did not must have been able to identify a preferred answer to each one of the 20 questions, whether that was the correct answer or not. However, only one of those 11 students had also selected exactly one answer to every question within the two other tests. In the third cohort, there were only four students who selected exactly one answer to every question across all three tests.

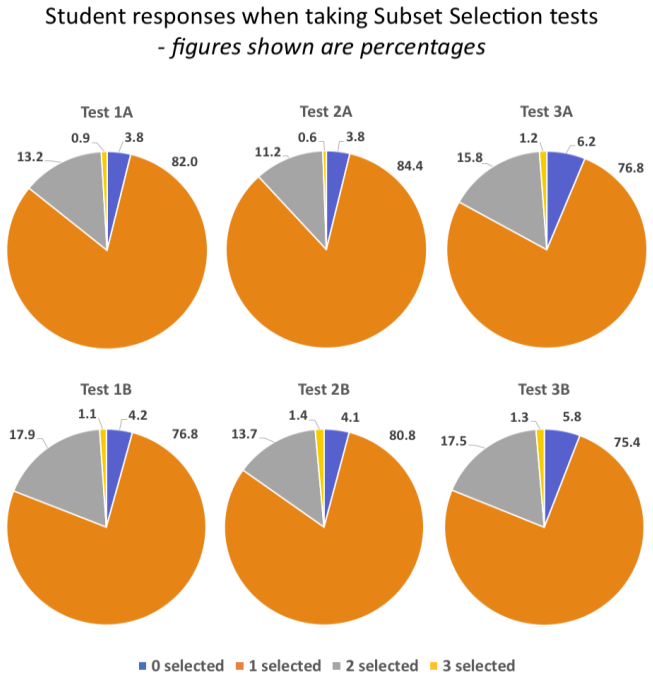

Figure 5. Student responses when taking the novel subset selection tests

Conclusion

The research revealed that the vast majority of students taking the novel subset selection tests (with the marking scheme described in section 3) do – at least sometimes – take advantage of the possibility of

selecting more than one option per question in an attempt to gain a partial mark rather than engaging in guesswork.

Less guesswork should imply more reliable test scores. Indeed, the findings of the empirical study indicate, although they do not show conclusively, that these novel subset selection tests do appear to yield more reliable test scores than traditional multiple-choice tests. In this sense, the findings are consistent with previous studies that focused on conventional subset selection tests (discussed in section 4.4).

The findings of the present study also show that these novel subset selection tests are less stressful for students than traditional multiple-choice tests. This is another important benefit. By contrast, many test takers find conventional subset selection tests relatively stressful (discussed in section 2).

It was a surprise to find that 42% of the participants involved in the empirical study felt that traditional tests are fairer. After subsequent discussions with some of the students, it appeared that this may have been partly due to some students feeling that the award of a 0.25 mark for each skipped question in the novel subset selection tests was unfairly generous. These students evidently had not fully appreciated that this is equal to the average mark obtained through blind guesswork in a traditional test.

A third benefit of these novel subset selection tests is that they do not yield inflated or (significantly) deflated marks with respect to traditional multiple-choice tests, as explained in section 3. This makes it easy for educators who currently use traditional multiple-choice tests to switch to using subset selection tests. (This is only true for the novel marking scheme described in section 3, it is not true for the conventional subset selection marking scheme described in section 2.3.) A slight deflation in the marks was noted in the empirical study, presumably due to some students occasionally 'playing safe' by ticking their second-choice option in addition to the correct option, thereby losing half a mark. This effectively introduces an element of confidence assessment into the tests, which may or may not be seen as a desirable feature.

A fourth benefit of subset selection tests in general, as compared with traditional multiple-choice tests, is that the responses they yield provide better feedback regarding, for example, whether there are any particular distracters that are deemed possibly correct by either a very high or a very low proportion of the students. This kind of feedback is helpful in two ways; it may shed light on topics that were not generally well understood by the students, and/or it may shed light on poorly designed test questions.

Overall, the findings of this study are very encouraging. Subset selection tests with the novel marking scheme described in section 3 are now in routine use within several courses at the authors' university. Some bespoke software has been developed to support these tests.

Finally, as is the case for any novel form of multiple-choice assessment, it is very important to give uninitiated test takers a clear explanation of the test format and the marking scheme, plus practice tests in advance of their first real test.

References

Akeroyd, F. M. (1982) Progress in multiple-choice scoring methods 1977-81. *Journal of Further and Higher Education*, 6(3), 87-90

Ávila, C., & Torrubia, R. (2004) Personality, expectations, and response strategies in multiple-choice question examinations in university students: a test of Gray's hypotheses. *European Journal of Personality*, 18(1), 45-59

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997) A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1), 65-88

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009) Does correction for guessing reduce students' performance on multiple‐choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, 34(1), 1-15

Bradbard, D. A., Parker D. F. & Stone, G. L. (2004) An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, 2(1), 11-26

Bush, M. (2001) A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2), 157-163

Bush, M. (2014) Multiple-Choice Tests Without the Guesswork at TEDxLondonSouthBankU [Video]. Available at https://www.youtube.com/watch?v=ACB2B2EdiXs (Last accessed 25-Jun-2018)

Caines, J., L. Bridglall, B. & Chatterji, M. (2014) Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education*, 22(1), 5-18

Diamond, J., & Evans, W. (1973) The correction for guessing. *Review of Educational Research*, 181-191

Dressel, P. L. & Schmid, P. (1953) Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 13, 574-595

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, 2(1), 79-96

Holmes, P. (2002) Multiple evaluation versus multiple choice as testing paradigm: feasibility, reliability and validity in practice. ISBN: 90-365-1757-5, Universiteit Twente PhD thesis available at http://www.ub.utwente.nl/webdocs/to/1/t0000017.pdf (Last accessed 25-Jun-2018)

Jaradat, D. & Sawaged, S. (1986) The subset selection technique for multiple-choice tests: an empirical inquiry. *Journal of Educational Measurement*, 23(4), 369-376

Jennings, S. & Bush, M. (2006) A comparison of conventional and liberal (free-choice) multiple-choice tests. *Practical Assessment, Research & Evaluation*, 11(8)

Karandikar, R. L. (2010) On multiple choice tests and negative marking. *Current Science*, 99(8), 1042-1045

Lance, C. E., Butts, M. M. & Michels, L.C. (2006) The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2), 202-220

Rippey, R. (1968) Probabilistic testing. *Journal of Educational Measurement*, 5(3), 211-215

Suen, H. K. & French, J. L. (2003) A history of the development of psychological and educational testing. Handbook of psychological and educational assessment of children, 3-23

Tavakol, M. & Dennick, R. (2011) Making sense of Cronbach's alpha. *International journal of medical education*, 2, 53.

Wood, R. (1991) Assessment and testing: a survey of research, Cambridge University Press, Cambridge

Zapechelnyuk, A. (2015) An axiomatization of multiple-choice test scoring. *Economics Letters*, 132, 24-27.

Citation:

Otoyo, Lucia, & Bush, Martin. (2018). Addressing the Shortcomings of Traditional Multiple-Choice Tests: Subset Selection Without Mark Deductions. *Practical Assessment, Research & Evaluation*, 23(18). Available online: http://pareonline.net/getvn.asp?v=23&n=18

Corresponding Author

Lucia Otoyo
Lecturer of Computer Science
London South Bank University
03 Borough Road, London, SE1 0AA

email: lucia.otoyo [at] lsbu.ac.uk