# Reducing the need for guesswork in multiple-choice tests

Martin E Bush

Department of Informatics

London South Bank University

UK

**Abstract**

The humble multiple-choice test is very widely used within education at all levels, but its susceptibility to guesswork makes it a sub-optimal assessment tool. The reliability of a multiple-choice test is partly governed by the number of items it contains, however longer tests are more time consuming to take, and for some subject areas it can be very hard to create new test items that are sufficiently distinct from previously used items. A number of more sophisticated multiple-choice test formats have been proposed dating back at least 60 years, many of which offer significantly improved test reliability. This paper offers a new way of comparing these alternative test formats, by modelling each one in terms of the range of possible test taker responses it enables. Looking at the test formats in this way leads to the realisation that the need for guesswork is reduced when test takers are given more freedom to express their beliefs. Indeed, guesswork is eliminated entirely when test takers are able to partially order the answer options within each test item. The paper aims to strengthen the argument for using more sophisticated multiple-choice test formats, especially for high-stakes summative assessment.

**Introduction**

A conventional multiple-choice test consists of a collection of questions or incomplete statements, often referred to as *stems*, each having a number of associated *options*. A stem together with its options is often referred to as an *item*. One of the options within each item is the right answer while the others are all wrong answers, often referred to as *distractors*.

This paper discusses the relative merits of eight inter-related multiple-choice test formats in which the same collection of items can be reused without modification. The traditional multiple-choice test format requires test takers to select one option only per item. In each of the other seven test formats the test takers are presented with a wider range of possible responses, giving them greater freedom to express their true belief in relation to each item.

The content of this paper should be of interest to all designers of multiple-choice tests, especially in the context of high-stakes summative assessment where the reliability of test scores is of major concern. Most of the test formats discussed have been described previously in the literature, but conflicts of terminology can easily lead to confusion. In this paper each test format is explained in a consistent way.

**Concepts & Terminology**

Every multiple-choice test has a *format* and an associated *marking scheme*. The format of a multiple-choice test corresponds to the *instructions for test takers*, while the marking scheme can be thought as the *instructions for test markers*.

A test taker faced with a given item will have a certain belief (*B*) in relation to its options. For the purposes of this paper, *B* is defined as a partial ordering of the options that reflects the relative confidence the test taker has in each option being correct. For example, *B* = <{*a*}, {*b*, *c*, *d*}> indicates greatest confidence in option *a* but no preference amongst the other three.

Each test format has an associated set of permitted responses, but the nature of those responses is not the same for every test format (as will become clear). Each response implies a certain belief on the part of the test taker, therefore it is convenient to define the *vocabulary* (*V*) of each test format as the set of implied beliefs corresponding to the permitted responses. For example, if a test taker responds to a given item within a traditional multiple-choice test by selecting option *b* for that item this implies that *B* = <{*b*}, {*a*, *c*, *d*}>.

In a traditional test there are only five permitted responses, so the vocabulary of the traditional test format is as follows:

*Vtrad* = { <{*a*}, {*b*, *c*, *d*}>, <{*b*}, {*a*, *c*, *d*}>, <{*c*}, {*a*, *b*, *d*}>, <{*d*}, {*a, b*, *c*}>, <{*a*, *b*, *c*, *d*}> }

The last element corresponds to a *skipped* question, where no option is selected.

Whenever test takers have equal confidence in all of the options (*B* = <{*a*, *b*, *c*, *d*}>) but they make a response which implies something other than this, that is *pure* guesswork. Negative marking involves the deduction of marks for incorrect selections to discourage test takers from engaging in pure guesswork. The number of marks deducted is normally chosen such that pure guesswork has no impact on the expected (mean) score. In other words, the positive marks due to lucky pure guesses and the negative marks due to unlucky pure guesses will cancel each other out on average. Tests that have this property are referred to as being *pure-guess-neutral* within this paper.

*Informed* guesswork occurs whenever test takers have *unequal* confidence in all of the options but they are unable to express their true belief and instead they must choose arbitrarily between two or more responses that coincide equally closely with their true belief. For example, a test taker with true belief <{*a*, *b*}, {*c*, *d*}> for a given item would be forced to choose arbitrarily between options *a* and *b* in a traditional test, corresponding to the implied beliefs <{*a*}, {*b*, *c*, *d*}> and <{*b*}, {*a*, *c*, *d*}> respectively.

The term “informed guesswork” gives the impression that the test takers have some partial knowledge with respect to a particular item, but this is not necessarily true. They may in fact be *misinformed* [Bradbard, Parker & Stone, 2004].

In the case of a *4-choice* test, in which each item contains four options, any randomly selected option is three times more likely to be incorrect than it is to be correct. Therefore, awarding +3 for every correct selection and -1 for every incorrect selection is a pure-guess-neutral marking scheme because it ensures that test takers have nothing to gain or lose, on average, by engaging in pure guesswork.

**Eight alternative multiple-choice test formats**

Numerous alternatives to the traditional multiple-choice test format have been proposed in the literature dating back at least 60 years [Dressel & Schmid, 1953], with varying levels of take-up. Some are quite well known, but others much less so. The following eight distinct test formats are discussed within this paper:

1. *traditional*,
2. *subset selection*,
3. *distractor selection*,
4. *strict ordering*,
5. *repeated selection*,
6. *repeated distractor selection*,
7. *partial ordering*,
8. *repeated subset selection*.

Several of these test formats are also known by other names. The discussion that follows assumes 4-choice tests throughout. Figure 1 indicates some of the inter-relationships between the eight test formats, and shows the number of possible responses per question for the more straightforward ones. This relates to the issue of *freedom of expression*, which is the subject of the next section. The exact number of possible responses for the more complex test formats is not easy to determine, hence the question marks in Figure 1.



Figure 1. Eight alternative multiple-choice test formats and their inter-relationships.

The three *repeated* test formats (shown shaded in Figure 1) naturally lend themselves to computer-based testing, while the other five may be presented either as “paper-and-pencil” tests or computer-based tests.

In the discussion that follows a pure-guess-neutral marking scheme has been chosen for each test format such that the mark per item ranges from +3 down to –3 in each case.

***Traditional***

For traditional multiple-choice tests the instructions for test takers can be paraphrased as follows:

Select the option that seems to you the most likely to be correct for each question.

This explains succinctly what the test takers are required to do, but when negative marking is used it is important to explain the marking scheme as well so that the test takers can see that they will risk losing marks if they engage in guesswork:

Select the option that seems to you the most likely to be correct for each question. Three marks will be awarded for every correct selection, minus one for every incorrect selection.

Negative marking is widely used in conjunction with the traditional test format. This is sometimes referred to as *correction for guessing* or *formula scoring* [Frary, 1988].

***Subset selection***

In a subset selection test the instructions for test takers can be paraphrased as follows, where once again the first sentence explains the test format and the second sentence explains the marking scheme:

Select the option(s) that seem to you the most likely to be correct for each question; you may select up to three options per question. Three marks will be awarded for every correct selection, minus one for every incorrect selection.

Subset selection tests are a generalisation of traditional multiple-choice tests in the sense that they offer test takers a wider vocabulary of possible responses. In other words, the vocabulary of traditional multiple-choice tests *Vtrad* is fully contained within the vocabulary of subset selection tests *Vsub*. The relationship can be expressed succinctly using mathematical set notation as follows: *Vtrad* ⊂ *Vsub*. This means that test takers who are faced with a subset selection test may choose to regard the test as being a traditional multiple-choice test with negative marking.

Subset selection tests appear to have been first proposed by Dressel and Schmid [Dressel & Schmid, 1953], who referred to them as *free-choice* tests. They have also been referred to as *liberal* tests [Bush, 2001] as well as *subset selection* tests (see e.g. [Jaradat & Sawaged, 1986]). Akeroyd’s *dual response* system [Akeroyd, 1982] is somewhat similar, although it does not use negative marking.

***Distractor selection***

In a distractor selection test the instructions for test takers can be paraphrased as follows:

Select the option(s) that you think correspond to *wrong* answers; you may select up to three options per question. One mark will be awarded for every wrong answer correctly identified, three marks will be deducted whenever a *right* answer is selected.

This is the inverse of subset selection. In a subset selection test the test takers are required to select the option(s) they believe is or are the most plausible, which implies identification and exclusion of the options they believe correspond to wrong answers. By contrast, in distractor selection the test takers are asked to select the distractors. The vocabulary of responses for distractor selection is the same as the vocabulary for subset selection: *Vdis* = *Vsub* .

Distractor selection has been referred to in the literature as the *Coombs Elimination Procedure* [Coombs, Milholland & Womer, 1956] and *Elimination Testing* [Ben-Simon, Budescu & Nevo, 1997]. This is probably the most widely used and written about type of pure-guess-neutral multiple-choice test after traditional multiple-choice tests with negative marking.

***Strict ordering***

In a strict ordering test the instructions for test takers can be paraphrased as follows:

Answer each question by ordering the options according to how likely each one seems to you to be the right answer, where ‘1’ indicates most likely and ‘4’ indicates least likely. Your mark will be +3, +1, –1 or –3 depending on whether you ranked the correct answer 1st, 2nd, 3rd or 4th (respectively).

The marking scheme described here is not the only possible marking scheme. For a pure-guess-neutral multiple-choice test the only requirement is to use a sequence of numbers for the marks that add up to zero. For example, +3, +2, +1 and –6 would be sufficient. This would reward partial knowledge more generously, since if a test taker could correctly identify one distractor then they would be guaranteed a positive mark.

This test format has been referred to as *complete ordering* in the literature [Ben-Simon, Budescu & Nevo, 1997]. Its vocabulary *Vstr* consists of the various permutations of the four options together with *skip* (no response).

In a traditional multiple-choice test the test takers can only select their first choice option; they may have a second and even a third choice in mind, but they are unable to express this. In that sense, it is tempting to consider strict ordering to be a generalisation of traditional multiple-choice tests because if their 1st choice answer is wrong then their 2nd choice answer will be looked at, and if that is also wrong then their 3rd choice answer will be looked at. However, in order to be a true generalisation of traditional multiple-choice tests its vocabulary would need to include the five traditional multiple-choice test responses, but those are not strict orderings.

***Repeated selection***

If a strict ordering test were to be delivered via a computer-based system this raises the possibility of enabling the test takers to submit each selection in turn in a repeated manner rather than having to rank all of the options prior to submission. This is the essence of a repeated selection test. The test takers are now spared the task of having to decide on an ordering for any remaining options after they have successfully identified the correct one.

Furthermore, a computer-based system can provide an additional mid-question *give up* option for test takers, rather like *skip*, to use when they want to give up trying to answer any particular question. For example, a test taker might choose to give up if their first selection was incorrect and the remaining three options appeared to them to be equally plausible. Whenever *give up* is invoked, the test can be marked as though it were a partial ordering test (described below).

The instructions for test takers can be paraphrased as follows:

Answer each question by first selecting the option that you think seems most likely to be the right answer, or you may choose to give up. If your first selection is incorrect, you may either make a second selection or give up at that point. If your second selection is incorrect, you may either make a final selection between the two remaining options or give up at that point.

These instructions might seem relatively complicated and off-putting for some test takers. Fortunately, having a computer-based system raises the possibility of making the instructions much more user friendly by only presenting relatively simple, contextual instructions as and when appropriate. The first instruction could simply say:

Select the option that you think most closely corresponds to the right answer, or give up.

…and then the second and third instructions (if required) could each say:

That was incorrect. Select another option, or give up.

Repeated selection tests have been referred to as *answer-until-correct* tests [Wood, 1991], [Persky & Pollack, 2008] and the *Immediate Feedback Assessment Technique* [Epstein et al, 2010]. The latter are associated with the use of “scratch off” answer sheets rather than computer-based testing.

The inclusion of a mid-question *give up* option makes the vocabulary of repeated selection tests *Vrep* rather large and complex, since it encompasses *Vstr* and*Vtrad* and other implied beliefs such as <{*a*}, {*b*}, {*c*, *d*}>. See Section 4 for a fuller discussion.

***Repeated distractor selection***

If a distractor selection test were to be delivered via a computer-based system, the test takers would not necessarily need to select all of the distractors within an item at once. Instead, the options could be selected one at a time, either until no more distractors can be confidently identified or the correct answer is wrongly identified as being a distractor.

In a sense this is the inverse of repeated selection, since here the test takers begin by selecting the option they believe is the *least* likely, rather than the *most* likely, to be the correct one. Their vocabularies are quite different though, as discussed in Section 4.

***Partial ordering***

In a partial ordering test the instructions for test takers can be paraphrased as follows:

Answer each question by ordering the options according to how likely each one seems to you to be the right answer, where ‘1’ indicates most likely and ‘4’ indicates least likely. You may assign an equal rank to any of the options, so that your ranking could be any one of the following: (1–2–3–4), (1–1–3–4), (1–2–2–4), (1–1–3–3), (1–1–1–4), etc. Your mark will be in the range +3 down to –3, and will depend on the rankings you have assigned and which option is correct.

This test format has been referred to as *partial ordering* in the literature [Ben-Simon, Budescu & Nevo, 1997]. It is a generalisation of both strict ordering and subset selection since *Vstr* ⊂ *Vpar* and *Vsub* ⊂ *Vpar* .

The marking scheme is based on the one given above for strict ordering: “Your mark will be +3, +1, –1 or –3 depending on whether you ranked the correct answer 1st, 2nd, 3rd or 4th (respectively)”. If the right answer is one of two or three options ranked equally, the mark awarded is the expected mark if strict ranking had been enforced (in which case the test takers would have had to make a guess). For example, if the right answer was ranked equal first with one distractor then the mark awarded would be the average of +3 and +1, which is +2. If the right answer was ranked equal first with two distractors, the mark awarded would be the average of +3, +1 and –1, which is +1. If the right answer was ranked equal last (i.e. equal 3rd) with one distractor, the mark awarded would be the average of –1 and –3, which is –2.

***Repeated subset selection***

If a partial ordering test were to be delivered via a computer-based system, the test takers could initially treat each item as though it was a subset selection test. If their first choice of option(s) does not include the right answer they could then be given a second chance, and then possibly also a third chance. A mid-question *give up* option could be included, and instructions for test takers can be contextual. If a test taker gives up mid-question this means that the remaining options are ranked equally in last place, so the marking scheme suggested above for partial ordering can also be used here.

**Freedom of expression**

The freedom of expression that a test format enables is partly determined by how many different responses it makes available for the test takers to choose between, as indicated in Figure 1, but this is not the whole story.

In a ***traditional***multiple-choicetest (with or without negative marking) the test takers must choose one of five possible responses per question: *a*, *b*, *c*, *d* or *skip*.

In a ***subset selection***test the test takers must effectively choose between 16 possible responses per question, since each option may or may not be selected. (Selecting all options is equivalent to selecting none, which is effectively *skip*.)

In a ***distractor selection***test the test takers must also choose between 16 possible responses per question, since each option may or may not be selected. (Again, selecting all options is equivalent to selecting none.)

In a ***strict*** ***ordering***test the test takers must choose between 25 possible responses per question, since there are 24 possible permutations of the options, with *skip* being the 25th response. (Skip is included here for pragmatic reasons; it is not a permutation of the options, but test takers may decline to answer any particular question.) With over 50% more responses to choose from as compared with subset selection and distractor selection, it could be argued that strict ordering offers greater freedom of expression. This would be misleading however, because the only response that is contained within both of their vocabularies is *skip*. Strict ordering forces test takers to engage in some informed guesswork whenever they can eliminate one or more distractors but not decide on a preferential ordering amongst the remaining options. This may be a common scenario, which subset selection and distractor selection both cater for.

In a ***repeated selection***test the test takers *may end up choosing* one of the 24 permutations, as in a strict ordering test, but they could hit upon the right answer at their first, second or third attempt, or give up after their first or second attempt. The response to each question will therefore depend on the combination of the test taker’s belief and which option is correct. For example, in the case of *B* = <{*a*}, {*b*}, {*c*}, {*d*}> the test taker’s response will differ depending on whether *a* is correct, *b* is correct, or one of *c* or *d* is correct. The vocabulary of this test format encompasses both *Vstr* and*Vtrad*  and it also caters for more complex beliefs such as <{*a*}, {*b*}, {*c*, *d*}>, which could correspond either to incorrect selection of *a* followed by correct selection of *b*, or incorrect selection of *a* followed by incorrect selection of *b* followed by *give up*. (In the first case the mark would be +1, in the second case it would be –2; see above for an explanation of the marking scheme for partial ordering.)

In a ***repeated distractor selection***test a test taker’s responses will also depend on the combination of their beliefs and which options are correct, as for repeated selection. Although this test format is in some sense the inverse of repeated selection, their vocabularies are quite different. The test takers start by selecting the option they believe is the *least* likely to be the correct one, and they may stop trying to select distractors at any point. Hence the vocabulary for this test format *Vrepdis* consists of implied beliefs in which the least likely options are ordered but not necessarily the most likely. For example, it includes <{*a, b, c*}, {*d*}> (corresponding to selection of *d* only) and <{*a, b*}*,* {*c*}, {*d*}> (corresponding to selection of *d* followed by *c*), but not <{*a*}, {*b*}, {*c*, *d*}>. It is a conjecture only at this point that *Vrep* and *Vrepdis* contain the same number of responses, although not the same responses.

In a ***partial*** ***ordering*** test the test takers are able to fully express their relative confidence in all of the options, since *Vpar* is the set of partial orderings of the four options.

In a ***repeated*** ***subset selection*** test the test takers also have complete freedom to express their relative confidence in all of the options, as in a partial ordering test. So the vocabularies are the same, but – as for the other repeated test formats – a test taker’s response to each question will in fact depend on the combination of their belief and which option is correct.

In terms of freedom of expression, the key point is that if test takers are restricted in their ability to express what they truly believe then this may compel them to engage in some informed guesswork. Partial ordering and repeated subset selection tests are the clear winners in this regard, since they eliminate the need for guesswork entirely.

It is possible to add a further dimension of belief expression to each one of the eight test formats described so far by requiring test takers to declare a *level of confidence* in their response to each item. For example, in an otherwise traditional test format the test takers could be asked to associate one of three confidence levels – *low*, *medium* or *high* – with each of their chosen options, and their score per item (which could be positive or negative) would then reflect not only whether they have answered correctly but also their level of confidence in their chosen option [Gardner-Medwin, 2006]. Given a suitable marking scheme, such tests can also be pure-guess-neutral.

In *probabilistic* tests [Rippey, 1968] and *multiple-evaluation* tests [Holmes, 2002] the test takers are required to associate a percentage “personal probability” to *every* option within each item, not just their most favoured option. These tests are based on the related notions of *admissible probability measurement procedures* and *symmetric reproducing scoring systems* which ensure “that any student, at whatever level of knowledge or skill, can maximize his expected score if and only if he honestly reflects his degree-of-belief probabilities” [Shuford, Albert & Massengill, 1966 – page 125].

**Suitability for given scenarios**

Multiple-choice tests are used within many different scenarios. They can be summative or formative, one-off or one of a series, paper-and-pencil or computer-based, designed for children or designed for adults. All of these factors should be taken into consideration when choosing a test format and an associated marking scheme. Whenever an unfamiliar test format or marking scheme is used it is important to spend time explaining the rules and any implications to the test takers in advance. Ideally they should also be given at least one practice test prior to any summative assessment.

***Traditional***

Traditional multiple-choicetests without negative marking may be the best choice for low-stakes, one-off summative tests because of their simplicity and familiarity. The use of negative marking may discourage pure guesswork, but some informed guesswork is to be expected. Without negative marking, traditional multiple-choice tests are arguably even worse than open response tests from the point of view of reliability [Azalov, Azalov & Zlatarova 2004] unless the test takers are incentivized in some other way to skip questions rather than make pure guesses. It is not necessary to use negative marking for this; for 4-choice tests it can be achieved by awarding one quarter of the credit given for selecting a right answer whenever a question is skipped [Akeroyd, 1982].

***Subset selection***

Subset selection is a very attractive proposition for summative tests. This test format has been used routinely since the late 1990s by the author at his university [Bush, 2001]. Subset selection tests reward partial knowledge not only more reliably but also more generously than traditional tests do when the marking scheme is as described earlier in this paper, because a test taker who can successfully identify two distractors within an item will get a *guaranteed* mark of 2 out of 3 versus an *expected* mark of 1.5 in a traditional test [Jennings & Bush, 2006].

***Distractor selection***

Distractor selection is another attractive proposition for summative tests. It is equivalent to subset selection in terms of freedom of expression, but there is a subtle difference between these two test formats, due to the fact that individuals may sometimes reach different conclusions when asked to identify a subset of plausible answers from a larger set of candidates through a process of explicit *inclusion* (as in the case of subset selection) versus a process of explicit *exclusion* (as in distractor selection) [Yaniv & Schul, 1997]. In other words, there could be one or more options within an item that a given test taker might not identify as being plausible answers and yet that same test taker might not identify them as being implausible answers either. Distractor selection tests may be more effective than subset selection tests at reducing the occurrence of pure guesswork [Warwick, Bush & Jennings, 2010].

One issue that came to light during a trial of distractor selection by the author at his university is that of double negatives, which arises whenever a item has a “not” in the question. For example, consider a question such as “Which one of the following should *not* be considered a symptom of … ”; in a distractor selection test the test takers must select the options which appear *not* to be the right answer to this question, which can be confusing. This problem can be overcome by re-phrasing the question and instructions as follows: “Only one of the following should *not* be considered a symptom of … ; identify as many as you can of the options that correspond to valid symptoms.”

Promising results have been obtained recently by researchers using distractor selection tests with five options per question [Bond et al, 2013]. Their test takers (university life science students) reported feeling less stressed when faced with a distractor selection test versus a traditional test. They also reported that they were generally not distracted by thinking about the best tactics for scoring high, and they generally agreed with the assertion that distractor selection tests helped to improve their critical thinking skills.

***Strict ordering***

Strict ordering is easy to understand, and it may seem attractive at first sight, but it is problematic. In a trial by the author at his university it was found not to work well when given as a paper-and-pencil test because many of the test takers just indicated their 1st choice options, leaving the remaining three unranked, as if they were taking a traditional multiple-choice test. Sometimes they indicated their 1st and 2nd choice options but left the remaining two unranked. This begs the question of what mark to award whenever a test taker has failed to assign a rank to a correct answer. This was dealt with by treating the test as if it was a partial ordering test for the purpose of scoring. If a strict ordering test were to be delivered via a computer-based system then the test takers could be forced to rank all of the options within every item, but in that situation a repeated selection test would be preferable.

***Repeated selection***

Repeated selection tests are now in use for formative assessment at the author’s university using a newly developed web-based system (“QuizSlides”). The test takers engage with each question in turn in a very interactive manner, and they find out how well they are doing and also what the right answer is to each question as they progress through the test. The system uses a pure-guess-neutral marking scheme; the mark per question is +3, +1, –1 or –3 depending on whether the correct option was selected at the 1st attempt, 2nd attempt, 3rd attempt or three wrong options were selected (respectively). test takers can choose to give up straight away, in which case they are awarded 0 marks, or if they give up after selecting one or two wrong options they are then awarded –1 or –2 respectively. The system immediately indicates what the right answer is whenever a test taker chooses to give up.

Since the test takers see how well they are doing as they go along it is debatable whether or not this test format is suitable for summative assessment, because any mid-test feelings of elation or despair may impair a test taker’s ability to concentrate on the remaining items. There is evidence [Epstein et al, 2010], [Maravić-Čisar et al, 2012], [Dibattista & Gosse, 2006] to suggest, however, that repeated selection is indeed suitable for summative assessment, and furthermore that the immediacy of feedback does significantly enhance learning.

***Repeated distractor selection***

Repeated distractor selection is quite easy to envisage, and it should be no harder to implement than repeated selection, but it offers no obvious advantages as compared with repeated selection. The author is not aware of any computer-based system that supports this test format.

***Partial ordering***

Partial ordering tests seem ideal for high-stakes summative assessment because they give test takers full freedom of expression. The marking scheme is relatively complex, however in the author’s experience of using this type of test within his university it is not as difficult for test takers to understand as it might at first appear. Marking the tests is also less difficult than it might at first appear. Even if some test takers use invalid rankings, such as (1–1–2–2), or they simply leave some of the options unranked, it is always possible – and not difficult – to translate their responses into the correct ranking. For example, (1–1–2–2) would be interpreted as (1–1–3–3). Given their complexity, however, they may not be suitable for children or for one-off tests. The marking procedure is not difficult, but it is not trivial either, so a computer-based solution would be preferable. The author is not aware of any computer-based system that supports this test format.

***Repeated subset selection***

Repeated subset selection tests are equally as good as partial ordering tests from the point of view of freedom of expression, but they differ in terms of whether or not the test takers get to see how well they are doing as they progress through a test. The author is not aware of any computer-based system that supports this test format.

**Closing remarks**

Traditional multiple-choice tests are by far the most common, but with the advent of more ubiquitous computer-based testing it is possible to envisage a different state of affairs. Repeated selection tests seem a natural choice for formative assessment, but for summative assessment the situation is less clear.

The simplicity and familiarity of traditional multiple-choice tests will no doubt guarantee their continued long-term popularity for summative assessment, but they are not ideal. Numerous past studies (e.g. [Jaradat & Sawaged, 1986], [Alnabhan 2002]) have shown that the use of some form of negative marking is effective at reducing pure guesswork, and that it therefore does lead to more reliable test scores. Not all studies have shown this, however. In particular, there is evidence to suggest that test takers are more inclined to engage in guesswork when they have relatively low expectations of gaining a high score [Bereby-Meyer et al, 2003], [Ávila & Torrubia, 2004].

Conversely, there is evidence to suggest that some test takers may be disadvantaged by negative marking due to their aversion to losing marks, and hence their reluctance to answer questions to which they are uncertain of the correct answer, even though they do have some partial knowledge [Rowley & Traub, 1977].

In the author’s experience, many test takers initially feel intuitively that any form of negative marking is fundamentally unfair, and that zero is a sufficient penalty in itself for wrong answers. Negative marking is an essential ingredient of any pure-guess-neutral marking scheme, and it is important for test takers to understand and feel comfortable with the reasons for its use.

Any of the eight test formats discussed within this paper can be presented in the form of a computer-based test, but the degree of interactivity is not the same in every case. The three *repeated* test formats require test takers to submit each individual selected option in turn, or – in the case of repeated subset selection – potentially a pair or triplet of options in which they have equal preference. This means that the test takers will (or at least may) discover the right answer to each question as they go along. This is certainly a desirable feature for formative tests, if not summative tests. By contrast, with the non-repetitive test formats the test takers may only submit their full response to each *item* in turn. In this case there can be no opportunities for second chances, and the test takers learn very little, if anything, about what the right answers are.

In the author’s experience, test takers need to be allowed a little more time per item when taking a non-traditional multiple-choice test as opposed to a traditional multiple-choice test. So if there is a fixed amount of time allocated for a test, the test takers can be presented with more items when the traditional test format is used. Since the reliability of any multiple-choice test is partly governed by how many items it contains there is something to be said in this respect for traditional multiple-choice tests, although the occurrence of any guesswork will reduce the test reliability. There is no simple answer to the question of how many items an multiple-choice test ought to contain [Burton, 2006].

Reusing items from an existing item bank can be risky and it is usually better to present test takers with new items that they could not possibly have seen before. In the case of a mathematics test it can be quite a simple job to produce many variations of a given item just by changing the numbers, but for other subject areas it is often a much harder task to create new items that are of high quality and sufficiently distinct from previously used items. Whenever the creation of new items is problematic, use of a non-traditional test format should be considered as a viable alternative to using a traditional test containing more items within the same allocated test time.

This paper has focused on the extent to which each of the eight test formats discussed enables test takers to express their true beliefs. Less guesswork implies more reliable test scores, which is especially important for high-stakes summative assessment. On this basis, *partial ordering* tests may be the best for high-stakes summative assessment because they eliminate the need for guesswork. Alternatively, if it is considered desirable for the test takers to see how well they are doing and thereby to learn what some or all of the right answers are as they go along, then *repeated subset selection* tests may be the best.

**Acknowledgements**

The author would like to thank Dennis Furey, Tony Gardner-Medwin, Sylvia Jennings, Lucia Otoyo and Jon Warwick for their thought-provoking input to this work.

**Notes on contributor**

Dr Martin Bush is a Principal Lecturer at London South Bank University, where he has spent over half his lifetime teaching a wide range of subjects in the general area of computing. He is also the founder of the online educational assessment company QuizSlides Ltd.

**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

Alnabhan, M. (2002) An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test, *Social Behaviour and Personality*, 30(7), 645-652

Azalov, P., Azalov, S. & Zlatarova, F. (2004) Work in progress – comparing assessment tools in computer science education: empirical analysis, *Proc. 34th ASEE/IEEE Frontiers in Education Conf*, 20-23 Oct 2004, Vol 2, F2G18-F2G19

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

Bereby-Meyer, Y., Meyer, J. & Budescu, D.V. (2003) Decision making under internal uncertainty: the case of mutiple-choice tests with different scoring rules, *Acta Psychologica,* 112(2), 207–220

Bond, A., Bodger, O., Skibinski, D., Jones, D., Restall, C., Dudley, E. & van Keulen, G. (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), e55956

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

Burton, R. (2006) Sampling knowledge and understanding: how long should a test be? *Assessment & Evaluation in Higher Education*, *31*(5), 569-582

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

Coombs, C. H., Milholland, J. E. & Womer, F. B. (1956) The assessment of partial knowledge, *Education and Psychological Measurement*, 16, 13-37

Dibattista, D. & Gosse, L. (2006) Test anxiety and the immediate feedback assessment technique, *The Journal of experimental education*, *74*(4), 311-328

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

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2010) Immediate feedback assessment technique promotes learning and corrects inaccurate first responses, *The Psychological Record*, *52*(2), 5

Frary, R. B. (1988) Formula scoring of multiple-choice tests, *Educational Measurement: Issues and Practice*, 7(2), 33-38

Gardner-Medwin, A. (2006) 12 Conﬁdence-based marking. *Innovative assessment in higher education*, 141-149

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 8-Sep-2011)

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)

Maravić-Čisar, S., Pinter, R., Radosav, D. & Čisar, P. (2012) Comparison of Examination Methods Based on Multiple-choice Questions Using Personal Computers and Paper-based Testing, *JITA-Journal of Information Technology and Applications (Banja Luka)-APEIRON*, *3*(1)

Persky A.M. & Pollack G.M. (2008) Using answer-until-correct examinations to provide immediate feedback to students in a pharmacokinetics course, *American Journal of Pharmaceutical Education*, 72 (4), article 83

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

Rowley, G. & Traub, R. (1977) Formula scoring, number-right scoring, and test-taking strategy, *Journal of Educational Measurement*, *14*(1), 15-22

Shuford, E. H., Albert, A. & Massengill, H. E. (1966) Admissible probability measurement procedures, *Psychometrika*, 31(2), 125-145

Warwick, J., Bush, M. & Jennings, S. (2010) Analysis and evaluation of liberal (free-choice) multiple-choice tests, *ITALICS*, 9(2)

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

Yaniv, I. & Schul, Y. (1997) Elimination and inclusion procedures in judgment, *Journal of Behavioral Decision Making*, 10, 211-220