Wrong is a relative concept: part marks for multiple-choice questions

Simon
University of Newcastle
simon@newcastle.edu.au

Abstract
Most of the literature concerning multiple-choice exam questions assumes that they are marked on a simple scale of 1 for a right answer and 0 for a wrong one. In this paper I present a case for multiple-choice questions that are worth more than one mark, for parity with other questions on the exam. I propose that with such questions, it is appropriate to award part marks for some of the wrong answers. I present some multiple-choice questions of this sort that have been used in the final exam for an introductory programming course, and analyse the students’ answers to confirm that the questions are valid assessment items.

Keywords: assessment, multiple choice, computing education, introductory programming

1 Introduction
Multiple-choice questions (MCQs) are used often, and probably increasingly, in computing education, typically because they save a great deal of time and effort in marking (Woodford & Bancroft 2005) – although much of the saving can be lost to the time and effort of constructing a good multiple-choice paper.

It appears to be the general belief that MCQs are flawed in that they can only test simple factual recall, but many authors (Lister 2000, Nicol 2007, Roberts 2006) argue that they can be used for higher purposes, with Lister, for example, suggesting that MCQs can test the first four levels of the Bloom Taxonomy (Bloom et al 1956).

In the final examination for an introductory programming course I have taken a small number of open-ended written questions and replaced them with multiple-choice questions. In two breaks with tradition, each of these questions is worth more than one mark, and some of the wrong answers are allocated part marks.

The rest of this paper explains why it was decided to replace the open questions with multiple choice and why it was decided to allocate multiple marks to each question with the possibility of part marks; examines the questions and their construction; analyses the questions in the light of the students’ responses; and concludes that the changes appear to have been reasonable ones.

2 Multiple-Choice Questions
Multiple-choice questions have been discussed at great length in the literature, and have their own terminology, which is summarised by authors such as Isaacs (1994) and Woodford and Bancroft (2005).

A question is called an item. The leading part of the question, preceding the multiple answers, is called the stem. The multiple answers are called options, the correct answer is called the key, and the incorrect answers are called distractors (or sometimes distracters).

There are many guidelines for writing good multiple-choice questions and for writing good multiple-choice tests (Gronlund 1982, Hansen & Dexter 1997, Isaacs 1994, Woodford & Bancroft 2005).

2.1 The traditional view of MCQs
Bloom’s Taxonomy of cognitive learning objectives (Bloom et al 1956) proposes a quasi-hierarchy of knowledge, comprehension, application, analysis, synthesis, and evaluation. There has been much debate about whether the categories do in fact form a hierarchy, what order they should take if they do (Anderson et al 2001), and whether in their current single form they are applicable to all areas of learning – Bloom himself apparently thought not (Anderson et al 2001). Even so, many educators take them as given, and frame their discussions accordingly.

The traditional view of MCQs is that they offer broad coverage of the material in a course, but at the expense of depth of coverage. Essentially, they can be used to test remembering (knowledge) and perhaps understanding (comprehension), but no more than those two levels of Bloom’s taxonomy.

Within these perceived limitations, MCQs can be used to test mastery of a topic, in what is known as criterion referenced assessment (have students grasped what they need to grasp in order to pass the course?). They can be used to rank students, in what is known as norm referenced assessment (how does each student score relative to the rest of the class?). And they can be used to find out what students do not know, in what is known as diagnostic assessment, with the intention either of remediating the students or of improving the course (Isaacs 1994).

There appears to be universal agreement that marking of multiple-choice examinations is easy. Even when the marking is not automated, it is probably easier than marking of open-ended questions. Further, analysis tools have been developed for distractors, for items, and for whole tests, so that once a test has been created and undertaken it is not difficult to validate it.
It appears to be widely acknowledged that while the use of MCQs reduces the time and effort spent marking, creation of a good MCQ test takes a great deal more time and effort than creation of a good open-ended test (Isaacs 1994, Lister 2000). For example, effort is required to ensure that each of the distractors is plausible to students who do not know the right answer, so that these students will not be able to guess the right answer simply by eliminating the implausible answers (Isaacs 1994, Hansen & Dexter 1997). One useful approach to choosing strong distractors is to base them on known student misunderstandings (Isaacs 1994).

2.2 Alternative views of MCQs

Notwithstanding the traditional views summarised in the preceding section, MCQs do not have to be quick and easy to answer, and are not limited to the assessment of recall and perhaps understanding.

One way to increase the cognitive load of answering a multiple-choice question is to make the stem longer and more involved. Questions with longer stems can take more time to read and absorb, but this can be offset by first describing a single involved scenario (called stimulus material), then asking several MCQs based on that same scenario (Isaacs 1994).

Even a short stem does not necessarily imply a speedy answer. In a mathematics exam, a stem such as ‘What is the value of the function \(5x^2 + x \sin(x)\) when \(x = 0\)?’ should not require a great deal of time from a well-prepared student, while ‘What is the integral with respect to \(x\) of the function \(5x^2 + x \sin(x)\)?’ would probably require significantly more time.

The previous example should also make it clear that MCQs can assess at higher Bloom levels than knowledge and comprehension. The second question is clearly one that requires application – unless, of course, the students have been shown exactly that example in class and are merely required to remember the right answer rather than to calculate it.

The questions presented above are incomplete, in that they are provided without their key and distractors. A full set of extremely implausible distractors could, of course, considerably reduce the cognitive load of the question. What makes a multiple-choice question work is the choice of suitable distractors, distractors that will indeed distract the less knowledgeable student by being plausible to that student.

Lister (2000, 2005) offers examples of non-trivial MCQs for a first programming course, with the understanding that each question will require substantial time from the students. These questions, he suggests, can assess at the lower four levels of Bloom’s Taxonomy.

Another variation from the simple memory-recall MCQ is the non-restricted item, a question with more than one correct answer, in which students are expected to select all of the correct options and none of the incorrect ones. Kolstad and Kolstad (1994) discuss the circumstances in which these questions might apply, and the various approaches to marking these questions, which can be quite intricate and clearly not a simple 0/1.

All of these variations offer ways of making MCQs more complex than the simple tests of recall that they are often taken to be, and suggest that the field is by no means yet fully explored.

3 A Different Perspective on MCQs

In a recent examination I have taken a small number of what were once open-ended questions and recast them as multiple-choice questions. In considering how to go about this, I made a number of decisions for which I found no clear precedent in the literature.

3.1 An MCQ can be worth more than 1 mark

In a test consisting entirely of MCQs, unless otherwise reported one assumes that the questions are worth 1 mark each, regardless of whether the questions are traditional recall questions or more involved ones such as those described by Lister (2000, 2005). In the latter case, it still appears that the questions are of comparable time or difficulty across the whole test, so it is reasonable to allocate the same mark to each of them. Lister (2000) does discuss the ‘brutality’ of the all-or-nothing approach of a complex MCQ, and concludes that this is akin to the brutality of a compiler either accepting or rejecting one’s code. This conclusion perhaps overlooks the fact that a compiler generally offers some assistance with removing errors, rather than considering them as final and penalising them accordingly.

In a test that mixes MCQs and open-ended written questions, the 0/1 marking scheme for the MCQs might no longer be suitable. With open-ended written questions, the mark generally reflects to some extent the time and effort required to answer the question. It would seem reasonable, then, to apply the same principle to the MCQs. If a written question requiring about 5 minutes of the student’s time is worth, say, 10 marks, surely a multiple-choice question that requires the same time should earn the same number of marks.

It follows that in some circumstances it might be contingent on the examiner to allocate more than one mark for certain MCQs.

3.2 Distractors can be allocated marks

In his essay The relativity of wrong Isaac Asimov (1988) observes that wrong is not a simple true/false concept. The Earth was once considered flat. Various scientists from several thousand to several hundred years ago postulated that it was actually spherical, and the postulate eventually became accepted fact. In recent decades, observations from space have led to the conclusion that it is in fact more of an oblate spheroid. But technically it is none of these, because of the perturbations of its mountain ranges, ocean trenches, and other smaller deviations from smoothness.

It is therefore wrong to say that the Earth is flat. It is also wrong to say that the Earth is round, but in some sense this would generally be considered less wrong than saying it is flat. In other words, wrong is a relative concept, and some wrong answers or ideas are more wrong than others.

Consider the following multiple-choice question:

The shape of the Earth is

(a) a plane
(b) an oblate spheroid
(c) a sphere  
(d) a torus.

Assuming that that we are prepared to ignore mountain ranges, ocean trenches, and other topographical anomalies, and that we remain blissfully unaware that the oblateness is marginally but measurably greater south of the equator than north of it, option b would be correct.

But if we chose to give some reward to a student who describes the Earth as a sphere in preference to a torus or a plane, we might decide to award half a mark to option c. This would be our recognition that not all wrong answers are wrong to the same degree.

Part of the reasoning for this notion stems from the observation (Isaacs 1994, Tew & Guzdial 2010) that student misconceptions make a good source of distractors in MCQs. If the same Earth-shape question were asked as an open-ended rather than a multiple choice question, a part mark would almost certainly be awarded to students who answered that it was spherical.

With open-ended written answers it is more or less obligatory to award part marks to answers that are partly but not fully correct: if the marks awarded are always all or nothing, it might be concluded that the marking scheme is not particularly appropriate.

Why, therefore, should it be unacceptable to award part marks for wrong answers simply because the form of the question has changed from open-ended to multiple-choice?

With these considerations in mind, I have used MCQs as part of a written exam, have allocated them a total mark commensurate with the time and effort required to answer them, and have allocated part marks for choosing some of the distractors.

It should be emphasised that this was not an experiment in setting a multiple-choice exam applying the new ideas of multiple marks and part marks. The experiment was to replace three highly specific written-answer questions in an exam with equivalent MCQs, for reasons that will be explained in the next section. However, the lessons learnt in this exercise can certainly be applied to exams that are fully multiple-choice.

3.3 Context

The course used as a context for this analysis is an introductory programming course offered at several campuses in Australia and overseas. Two offerings of the course have been analysed, one at an Australian campus and one at an overseas campus. There were 150 students in the two offerings, split fairly evenly between the campuses.

The programming language taught in this course is Visual Basic. Readers accustom to other C-based languages should not be alarmed at the lack of semicolons in the code provided in the questions.

4 The Questions

The BRACElet project (Whalley & Lister 2009) suggests that examinations include questions testing students’ ability to trace code, to read and explain code, and to write code. The code-explaining questions present students with a small piece of code and ask them to explain what the code does, not line by line but as an overall purpose. Many students do not score well on these questions.

When analysing the code-explaining questions, the project pays attention to the SOLO level (Lister et al. 2006, Whalley et al. 2006, Sheard et al. 2008) at which students answer. While the preferred response is relational, an explanation of the overall purpose of the code, many students give a multistructural answer, explaining the code line by line. Some BRACElet papers (Sheard et al. 2008) discuss the possibility that this is because despite ample coaching, students do not understand what kind of answer is being sought. Other BRACElet papers (Simon 2009) wonder whether students’ problems with code-explaining questions are due simply to their difficulties with written English expression. This might particularly be a problem for students whose first language is not English, whether at an overseas or at a local campus.

To address these possible confounds I decided to replace the open-ended code-explaining questions with comparable multiple-choice code-explaining questions. All of the multiple-choice options are expressed in English, and all are in the SOLO relational form, so when students cannot answer these questions correctly it is not because they cannot express their understanding correctly and not because they fail to appreciate what type of explanation is being sought; it is unequivocally because they do not understand the code.

While I initially imagined that it would be difficult to find plausible distractors for questions of this sort, when I turned back to earlier exams that had included similar questions in the open-ended written form I found a number of recurring wrong answers. For example, students forgetting that arrays indexes begin at zero tended to think that an iteration through an array was omitting the last element. I used these common wrong answers or their equivalent as the multiple-choice distractors, conforming with the recommendation (Isaacs 1994) that to be plausible, distractors should correspond to common student misunderstandings. However, because I really wanted to know what students thought was the purpose of each code segment, I added an open-ended option that permitted students to write what they thought the purpose was if they felt unable to choose any of the other options.

In previous exams for this course the thee code-explaining questions were each worth five marks, as a reasonable reflection of the time and effort it would take to answer them. Recasting the questions in multiple-choice format makes no appreciable difference to that time and effort, so when creating the exam I decided that the questions would be worth the same five marks, although at the time I imagined awarding full marks for the right answer and no marks for any of the wrong ones.

When marking began I was reminded of the brutality (Lister 2000) of all-or-nothing multiple-choice marking. In particular, I realised that while in previous offerings students would have been awarded some of the five marks for an answer that was in some way close to the right answer, students were now being awarded no marks at all for choosing that same answer from a list.

I therefore decided to award part marks for these questions. Specifically, a wrong distractor would be
awarded the same mark that the same answer would have earned in response to a comparable open-ended question. This mark was easy to determine, because I had in previous years set and marked comparable open-ended questions.

This approach has an obvious consequence for students who choose the ‘none of the above’ option and write their own explanation of the purpose of the code: such answers would be marked in exactly the same way as the equivalent open-ended question – except that a completely correct answer would not earn the entire maximum mark, because the student had failed to recognise the correct answer among the options.

The three questions are progressively more complex, as suggested in the BRACElet 2009.1 (Wellington) specification (Whalley & Lister 2009). The first question involves non-iterative code; the second, iterative code without control logic within the loop; and the third, iterative code with control logic within the loop.

The set of three questions was prefaced with the following instruction:

The next three questions are multiple choice, but please answer them in the 12-page answer booklet. If your answer to a question is one of a-d, you need only write the question number and the letter of the answer. If your answer is e, please write the question number and the letter e, then your own explanation.

4.1 Question 24 and marks

The first of the three questions tests only the application of assignment statements and simple arithmetic. However, the outcome of the code is unlikely to be self-evident to most novice programmers, who would therefore need to desk-check (hand-execute) the code to be sure of the answer. This observation contrasts with the expressed preference of Lister and Whalley (2009) that students determine the purpose of the code just by reading it, not by tracing it.

The number to the left of each option is the mark that was awarded to students choosing that option.

24. What is the purpose or outcome of the following piece of code?

\[
\begin{align*}
  b & = a + b \\
  a & = b - a \\
  b & = b - a
\end{align*}
\]

1 (a) to find the sum of the values of a and b
1 (b) to find the difference of the values of b and a
2 (c) to find the sum and the difference of the values of a and b
5 (d) to swap the values of a and b
? (e) something else – please write the purpose

4.2 Question 25 and marks

The second question tests the application of a simple counting loop that iterates through an array executing a simple arithmetic assignment statement.

25. What is the purpose or outcome of the following piece of code?

```
For i = 0 To dPayment.Length - 1
  dBalance = dBalance + dPayment(i)
Next
```

1 (a) to add a payment to a balance
0 (b) to count the payments
3 (c) to add all payments except the last to the balance
5 (d) to add all payments to the balance
? (e) something else – please write the purpose

4.3 Question 26 and marks

The third question tests the application of a simple counting loop, again iterating through an array, but this time involving a boolean decision.

26. What is the purpose or outcome of the following piece of code?

```
sOne = sTitle(0)
For i = 1 To iTitleLastIndex
  If sTitle(i).Length > sOne.Length Then
    sOne = sTitle(i)
  End If
Next
```

5 (a) to find the longest title in the array of titles
2 (b) to find the length of the longest title in the array of titles
2 (c) to move the longest title in the array of titles to the first place in the array
1 (d) to sort the array of titles according to title length
? (e) something else – please write the purpose

5 Analysis of the Questions

There are standard techniques for analysing MCQs by examining the students’ answers. In this section I describe three such techniques, apply each technique to these questions and, where appropriate, propose a modified form of the analysis to suit the novel aspects of these questions.

5.1 Item facility

The facility of an item or question is a measure of how easy that question is for students to answer, defined as the mean mark of all students for the question (Isaacs 1994). With traditional multiple-choice questions, this is simply the proportion of students who answer the question correctly, because correct answers score one mark and incorrect answers score zero.

Isaacs (1994) suggests that if the purpose of a test is to rank students, a facility of 40% to 60% is generally considered desirable, whereas if the purpose is to determine students’ mastery of the subject matter, the aim should be a facility of more than 80%. However, he
acknowledges that the true aim of a test is usually a hybrid of the two purposes.

I have measured two different facilities for each of these questions, as shown in Table 1. The ‘binary facility’ is the simple proportion of students who answered the question completely correctly, i.e. who chose the key as their answer; the ‘continuous facility’ is the average mark over all students for the question, and thus acknowledges the part marks that were awarded for some distractors including option e, the open-ended answer.

We see from Table 1 that none of the questions has a facility in the range suggested for mastery, but that the adoption of part marks for some wrong distractors does noticeably raise the facility of each question (in that the continuous facility of each question is higher than its binary facility). This is an inevitable consequence of the fact that some students who did not choose the key were nevertheless awarded some marks for the options they chose.

5.2 Item discrimination

The discrimination of an item is a measure of how well it distinguishes the good students from the poor students. From that definition alone it is clear that discrimination is of more interest in a test whose purpose is to rank students than in a test whose purpose is to assess students’ mastery of the subject matter. There are several approaches to the calculation of discrimination, varying, for example, on what proportion of students make up the ‘good’ and the ‘poor’ students. Gronlund (1982) divides the class into thirds according to their result on the whole test, and defines the discrimination of a question as the number in the upper third who answered the question correctly, minus the number in the lower third who answered it correctly, divided by the number in each third.

The greatest possible discrimination is 100%, achieved when all of the good students and none of the poor students answer the item correctly. The worst discrimination is –100%, achieved when none of the good students and all of the poor students answer it correctly.

As with item facility, we are interested in the ‘binary discrimination’ of each question, based on whether students chose the single correct option or one of the distractors, but also in a ‘continuous discrimination’, based on the average marks of students in the upper and lower thirds. We calculate the continuous discrimination by subtracting the average mark of the lower third of the class from the average mark of the upper third, and dividing the result by the maximum mark for the question.

The discrimination measures, shown in Table 2, indicate that all of the questions have positive discriminations, which is good. It also shows that for each question the continuous discrimination is lower than the binary discrimination. This is to be expected, as the award of part marks blurs the lines between those who can answer the question correctly and those who can’t.

5.3 Distractor usefulness

Mitkov and Ha (2003) estimate the usefulness of a distractor (also known as its effectiveness) by comparing the numbers of students in the upper and lower groups who selected that distractor. They suggest that a distractor can be good, attracting more students from the lower group than from the upper group; poor, attracting more students from the upper group than from the lower group; or not useful, attracting no students.

Table 3 shows, for each distractor from each question, the number of students from the upper and lower thirds who selected that distractor. It follows from these figures that all of the distractors are good: all of them distracted some students, and all distracted more students from the lower third of the class than from the upper third.

6 Discussion

In a final exam that also incorporates 20 traditional multiple-choice questions and a number of open-ended written questions, I have included three multiple-choice questions intended to discern whether students can correctly determine the purpose of a small segment of program code.

While the traditional multiple-choice questions were allocated one mark each (and thus 1% of the maximum possible mark for the exam), the three non-traditional questions were judged to require significantly more time and effort, and so were allocated more marks – in this case, five marks per question.

While the multiple-choice options included the correct answer and a number of plausible distractors, I believed that some students might not find any of the answers to their satisfaction, so I included a ‘none of the above’ option, and asked students choosing this option to furnish their own answers.

In keeping with the way that the same questions would have been marked if they were open-ended rather than

| Table 1: Facility of each question, both in 0/1 sense (binary) and in actual mark sense (continuous) |
|---------------------------------|-----|-----|-----|
| binary facility                 | Q24 | Q25 | Q26 |
| continuous facility             | 51% | 53% | 46% |
| continuous facility             | 61% | 75% | 62% |

| Table 2: Discrimination of each question in both binary and continuous senses |
|---------------------------------|-----|-----|-----|
| binary discrimination           | Q24 | Q25 | Q26 |
| continuous discrimination       | 37% | 33% | 51% |

| Table 3: Usefulness of each distractor from each question: numbers of students from the upper and lower thirds who selected it |
|---------------------------------|-----|-----|-----|
| fixed option (a)                | 0 : 5| 0 : 4| –   |
| fixed option (b)                | 1 : 2| 0 : 2| 0 : 3|
| fixed option (c)                | 1 : 9| 9 : 19| 3 : 9|
| fixed option (d)                | –   | –   | 1 : 12|
| open-ended option (e)           | 5 : 12| 1 : 4| 1 : 7|
With regard to the usefulness of distractors (Table 3), every one of the distractors was classified as good: none of them was rated poor or not useful. This is particularly gratifying as, when I set out to write these questions in multiple-choice form, it was by no means obvious that I would be able to devise any meaningful distractors.

6.1 Traditional question analysis
A number of the measures traditionally used to analyse multiple-choice questions have been applied to these three questions. With regard to facility (Table 1), each of the questions has a facility within the range suggested appropriate for ranking tests, while none of them has a facility within the range suggested appropriate for mastery tests. As the exam is a hybrid of these two, the facility of the questions appears appropriate.

With regard to discrimination (Table 2), each of the questions discriminates reasonably between students in the top third of the class overall and students in the bottom third.

With regard to the usefulness of distractors (Table 3), every one of the distractors was classified as good: none of them was rated poor or not useful. This is particularly gratifying as, when I set out to write these questions in multiple-choice form, it was by no means obvious that I would be able to devise any meaningful distractors.

6.2 Novel question analysis
The measures of item facility and item discrimination are usually predicated on the expectation that each question is worth a single mark, and each answer will be awarded either the full mark or zero. As these questions were worth multiple marks, and parts marks were possible, I have supplemented these traditional measures (which I now call binary facility and binary discrimination) with what I call continuous facility and continuous discrimination.

The continuous facility shows that these questions are easier than if they were marked all-or-nothing; that is, their continuous facility is higher than their binary facility. This is the expected consequence of allocating part marks to the questions, which will of course lead to a higher average mark.

The questions discriminate less clearly between higher- and lower-performing students with the continuous discrimination than with the traditional binary discrimination. Again this is in accord with expectations, as it replaces a stark black and white marking scheme with one that recognises shades of grey.

6.3 Conclusions
As some other authors have asserted, and in contrast with the traditional view, multiple-choice questions are not restricted to the testing of pure recall. They can also be used to test aspects of learning that are considered as higher-level, for example according to Bloom’s taxonomy.

It is possible and practical to allocate multiple marks to a single multiple-choice question. Where multiple-choice questions are mixed with open-ended questions in a single test, the allocation of multiple marks permits the mark for each question to reflect the time and effort required to answer that question.

In recognition that wrong is a relative concept, it is possible and practical to award part marks for selection of a wrong answer to a multiple choice question. Where the distractors clearly display differing degrees of wrongness, this is recommended as a way to distinguish students who are more or less on the right track from those who have no idea.

The decision to allocate part marks for certain distractors would normally go hand in hand with the decision to allocate multiple marks to multiple-choice questions. However, it could equally well be applied to single-mark questions, with fractional marks allocated and tallied by a spreadsheet or other software.

6.4 Implications and consequences
What are the consequences for educators of the ideas developed in this paper?

First, the ideas might lead some educators to give deeper consideration to the use of multiple-choice testing. Those who have always thought of a multiple-choice exam as a large number of one-mark questions testing factual recall might now appreciate that there is scope for using multiple-choice questions to examine other levels of knowledge, and that those questions need not be ‘short’ in the sense that they can be answered almost as quickly as they can be read. Programming educators who are newly considering the possibility of longer multiple-choice questions are strongly advised to read Raymond Lister’s ACE2005 paper (Lister 2005).

Does this different approach to MCQs defy the reasons for using MCQs in the first place? Not at all. It seems that the main reasons for using MCQs are consistency and ease of marking, and an MCQ with multiple marks and part marks for wrong answers can be marked automatically just as a traditional 0/1 MCQ can. Whether the answers are entered directly into a computer by students, scanned from a mark-sense answer sheet, or laboriously read by human markers, some software such as a spreadsheet is then used to allocate marks to each student. It is a trivial matter to apply in a spreadsheet the ideas presented in this paper.

One idea from this paper that does not lend itself to automated marking is the option e in each question: if you don’t believe any of the other options is right, write your own answer. But this option is by no means integral to the marking approach, and can be readily dispensed with. Readers are reminded that the option was provided in this exam solely because I wanted to know what students really thought the answer was—not exactly a standard goal in assessment!

Will the ideas presented here lead to lower failure rates? I hope not. I hope that whatever form of assessment is used in computing education, it is used well and wisely, and applied in such a way that the most capable students do well, the barely capable students pass, and the students who have not (yet) achieved capability fail. This should be no different when using multi-mark part-marked MCQs. One difference that might become evident, though, is a smoother gradation of marks across the range.

One consequence of the multiple-mark notion is that a fully multiple-choice exam need no longer be constrained to have all its questions of comparable ‘length’, that is, requiring comparable time and effort to answer them. It would be perfectly reasonable to construct an exam in which, say, the first 20 questions are traditional short
MCQs worth one mark each, the next eight questions are mid-length MCQs worth five marks each, and the final four are seriously long MCQs worth ten marks each – always bearing in mind that much of the effort saved in marking will almost certainly be spent instead on the creation of the questions and their options.

7 Future Work

If it proves feasible, I intend to conduct a quantitative comparison of students’ results under the written-answer approach and under the multiple-choice approach.

In line with the SOLO analysis carried out within the BRACElet project, it would be interesting to examine the answers written by students who chose option e. Is the proportion of multistructural answers reduced? If so, might it be possible to attribute this to the provision of several plausible answers, all in the relational form?

While my own exam for this introductory programming course has long consisted of a number of traditional 0/1 MCQs and a number of longer open-ended questions, consideration will now be given to replacing more of the longer questions with MCQs, and perhaps ultimately to making the exam fully multiple-choice.

8 Acknowledgements

I am grateful to the BRACElet project for the suggestion that programming examinations should include code-explaining questions, and for the copious analysis that its members have carried out on the worth of such questions for both assessment and research. I am also grateful to the anonymous reviewers of this paper: their suggestions have unquestionably improved the paper.

9 References


