Asymmetry in student achievement on multiple choice and constructed response items in reversible mathematics processes

Noname manuscript No.
(will be inserted by the editor)

Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes

Christopher J. Sangwin and Ian Jones

Abstract In this paper we report the results of an experiment designed to test the hypothesis that when faced with a question involving the inverse direction of a reversible mathematical process, students solve a multiple-choice version by verifying the answers presented to them by the direct method, not by undertaking the actual inverse calculation. Participants responded to an online test containing equivalent multiple-choice and constructed-response items in two reversible algebraic techniques: factor/expand and solve/verify. The findings supported this hypothesis: Overall scores were higher in the multiple-choice condition compared to the constructed-response condition, but this advantage was significantly greater for items concerning the inverse direction of reversible processes compared to those involving direct processes.

Keywords Assessment; algebra; calculus; symbolic manipulation

C. J. Sangwin, School of Mathematics, University of Edinburgh and I. Jones, Mathematics Education Centre, Loughborough University. E-mail: C. J. Sangwin@ed.ac.uk

1 Introduction

Summative assessment of students is a key part of education. In mathematics, assessments typically attempt to measure one or both of procedural knowledge and conceptual understanding (Rittle-Johnson & Siegler, 1999). Our focus here is on procedural knowledge, which has been defined as "the ability to execute action sequences to solve problems" (Rittle-Johnson, Siegler, & Alibali, 2001). High-stakes examinations around the world have been criticised for privileging procedural over conceptual items (e.g. Berube, 2004; Iannone & Simpson, 2012; Noyes, Wake, Drake, & Murphy, 2011). Part of the reason for this emphasis on procedural items is that they are relatively easy to produce and can be scored objectively (Swan & Burkhardt, 2012). As such, scoring reliabilities tend to be very high in mathematics compared to other subjects (Brooks, 2004).

This can tempt us to conclude that assessing procedural knowledge is straightforward and unproblematic, and perhaps compared to assessing conceptual understanding that is indeed the case (Bisson, Gilmore, Inglis, & Jones, 2016). However, high reliability scores do not necessarily indicate that items are valid (Wiliam, 2001). For example, ostensibly the same question presented in different formats (e.g. multiple choice versus constructed response) can produce different patterns of results across a sample of students (Martinez, 1999; Shepard, 2008). Moreover, the reversibility of many mathematical operations (e.g. 'expand the brackets' versus 'factorise') can result in examiners failing to assess what they intended to assess (Friedman, Bennett, Katz, & Berger, 1996), and not being aware that they have failed. This latter threat to validity is the focus of the research reported here.

We begin by considering and comparing two common question formats for procedural items, multiple choice (MC) and constructed response (CR). Our review of the literature leads to the hypothesis that the validity of MC items, but not CR items, is likely to be undermined by the reversible nature of common mathematical operations. We then define reversibility and provide examples of 'direct' and 'inverse' processes involved in many mathematical operations. Following this we present a study in which undergraduate students ($N = 116$) were administered procedural items involving reversible operations in both MC and CR formats. The pattern of results strongly indicates that what we define below as 'inverse' items did not perform validly when presented in a MC format. We conclude that presenting such items to students using a CR format would significantly improve validity.

1.1 Multiple choice (MC) and constructed response (CR) formats

Procedural MC items typically present a mathematical object, such as an equation, and an instruction to transform the object into a specified form. Here is an example of an item used in the study reported below.

Factorise: $64m^3 - 125$
1. $(8m - 5)(8m + 25)$
2. $(4m - 5)(16m^2 - 20m + 25)$

3. $(4m - 5)(4m^2 + 20m - 25)$
4. $(4m - 5)(16m^2 + 20m + 25)$

The equivalent CR item would contain the same question stem, 'Factorise: $64m^3 - 125$', but the answer options would be removed and replaced with a space to write the answer, or a text box if administered as a computer-based assessment. Removing the options from MC items in this way creates what are called stem-equivalent CR items (Friedman et al., 1996).

Shepard (2008) reviewed 16 studies that compared CR and MC item formats in a variety of disciplines, and reported that question format appears to have little effect on assessment outcomes for stem-equivalent items. She argued that such study designs add little useful information because authors

> carefully controlled for everything else, including content, cognitive process, and construct. The finding is essentially a tautology. Yes, if you strictly constrain multiple-choice and constructed response (*sic.*) items to be identical, predictably they measure the same thing. (Shepard, 2008, p. 605)

An instrument used by Friedman et al. (1996) consisted of algebra story problems of a classical type and, consistent with Shepard's review, they found no evidence of format effects between MC and CR problems. To look for proposed mechanisms they used a think aloud protocol to gather qualitative data.

> Similarities between formats occurred because subjects solved some CR and MC items using similar methods. A typical MC approach is to plug in the response options, looking for one that satisfies the constraints of the item stem. Surprisingly, subjects used this strategy with CR items as frequently as with MC items. Subjects appeared adept at estimating plausible answers to CR items and checking those answers against the demands of the item stem. In other words, subjects frequently generated their own values to plug in. (Friedman et al., 1996, p. 1)

One reason they found no format effect is that subjects were using a verification strategy for both CR and MC items. That is, and perhaps unexpectedly, subjects used a strategy commonly associated with the MC format to answer CR items with equal frequency.

A further review of comparisons between CR and MC formats was reported by Martinez (1999), who suggested that *"The similarity in what is measured by counterpart items of multiple-choice and CR formats is a mixed picture"*. For example, Bridgeman (1992) considered the extent to which CR versions of stem-equivalent MC items led to similar outcomes. The study involved items requiring numerical answers, allowing them to be machine scored in CR format. Bridgeman reported some differences between formats, *"when the multiple-choice options were not an accurate reflection of the errors actually made by students"*. Similarly, Kamps and van Lint (1975) undertook a controlled comparison of equivalent CR and MC tests in university mathematics and found a format effect. All students sat both tests, but the order in which they were administered (MC then CR, or CR then MC) was randomly allocated. The authors reported a moderate correlation ($r = .57$)

Direct	Inverse
Multiplication of numbers	Prime factoring of integers
Laws of Exponents	Laws of logarithms
Expanding brackets	Algebraic factoring
Single fraction	Partial fraction
Differentiation	Symbolic integration
Verify a solution	Solve an equation

Table 1 Reversible symbolic processes in elementary mathematics

between scores on the CR/MC formats, suggesting the CR and MC formats were not equivalent.

Other researchers have reported a gender effect on question format. For example, Hassmén and Hunt (1994), Mazzeo, Schmitt, and Bleistein (1993) and Livingston and Rupp (2004) found that achievement for males is higher than achievement for females when MC items are used. Goodwin, Ostrom, and Scott (2009) considered possible gender differences in the frequency of employing 'back substitution' as an informed guessing strategy on MC test items. However, they found no gender difference in performance on MC items that allow for back substitution strategies, even when controlling for possible confounds such as prior achievement in mathematics.

1.2 Reversible processes in mathematics

Goodwin et al. (2009) used the phrase 'back substitution' for the process of verifying whether a value is a solution to an equation. For example, values such as $x = -5$ and $x = 2$, are substituted into the equation $x^2+3x = 10$ to verify whether they are solutions. While Goodwin et al. used the term 'back substitution', they did not define it in detail. In this section we consider how 'back substitution', might be defined and operationalised. To do this we introduce the notion of 'reversible processes'.

Mathematics involves many symbolic manipulations that are reversible. The construct we wish to discuss is the more general notion of reversible symbolic processes in formal mathematics methods. For example, multiplying brackets is accompanied by the reverse process of factoring, as in $(x-1)(x+1) = x^2-1$, and the two written forms are said to be algebraically equivalent. We therefore consider factor/expand to be a reversible process. Examples of reversible processes from elementary algebra and calculus are listed in Table 1.

There are mathematical and educational aspects to the processes we have chosen to describe as reversible. The educational aspects are situated historically and culturally mediated, and we return to this below. First, we propose four hallmarks with which we can distinguish two directions, which we call 'direct' and 'inverse'.

1.2.1 Mathematical aspects of reversibility

Hallmark 1: Complexity. Added complexity does not qualitatively change direct processes, but can qualitatively change inverse processes. For example, when mul-

tiplying two polynomial terms the process does not significantly change in nature when the complexity of the terms changes, and multiplying many brackets is an inductive process. However, trying to find a factored form is qualitatively different. Factoring multi-variable polynomials is, in general, only taught in special cases, such as taking out a common factor or the difference of two squares or cubes.

Because of this added complexity, the inverse process is often taught as a number of separate methods for dealing with different cases. This is particularly marked in the case of the last reversible process in Table 1: verify/solve. One of the purposes of manipulating an equation into a standard form is to recognise which *type* of equation (e.g. linear, quadratic) we have and so guide which technique is needed to solve it. Hence, solve includes a very wide range of different techniques. Verifying and evaluating only requires the substitution of variables for values, and subsequent numerical computations. Techniques for solving equations can be mechanical, but identifying which algebraic moves are needed to solve even simple linear equations involves more decision making than verifying that a particular value is a solution.

Hallmark 2: Guess and check. Students are sometimes taught the inverse process by a "guess and check" method to reduce the inverse process back to the previously learned direct process. For example, when factoring a quadratic the integer factors of the constant term can be used to guide the guess and check. Symbolic integration often relies on an informed guess and check procedure.

When factoring a cubic, $p(x)$, one common contemporary approach is to guess a root a and verify that $p(a) = 0$. This information enables one factor to be taken, resulting in a quadratic problem remaining. Part of the didactic contract (Brousseau, 1997) with students is that examples encountered in tutorial problems (and high-stakes examinations) will be amenable to such techniques. In this case the integer factors of the constant term in the polynomial guide which values of a to choose in the first instance.

Hallmark 3: Confirmation. We would expect students to confirm their result when undertaking the inverse process by performing the direct process. We would not expect students to do the reverse. This is a natural consequence of Hallmark 2.

Hallmark 4: Computer Algebra Systems (CAS). CAS implement the student's algorithm (or something very close to it) for direct processes. However, CAS do not implement the inverse processes in the same way students are typically taught. Most inverse processes rely on techniques which were only developed from mathematical research undertaken in the late twentieth century specifically for CAS. For example, given an elementary expression[1], differentiation is a mechanical procedure with definite rules. These rules are extensible in the sense that while new functions require new rules, they extend what has already been learned. Integration is rather different. Indeed, constructing a definite algorithm for deciding whether a symbolic anti-derivative exists as an elementary expression, and if so computing it (i.e. symbolic integration), was only resolved comparatively recently, (e.g., Risch, 1969). This technique for integration is not taught, even to most university mathematics students. This is also true of factoring, see Davenport, Siret,

[1] An expression built up from addition, multiplication, and substitution from numbers, variables and the basic exponential, logarithmic and trigonometric functions. E.g. $\sin(x^2)$ or e^{-x^2}.

and Tournier (1993). Therefore, for the inverse direction there is a significant disconnect between what is actually taught and the general methods used by CAS, and we believe good educational reasons persist for this disconnect.

1.2.2 Educational aspects of reversibility

Despite these four hallmarks, students might be taught processes for performing particular inverse methods directly. For example, to factor a quadratic expression we could first complete the square and then take the difference of two squares as exemplified by the following.

$$x^2 - 6x + 5 = (x-3)^2 - 2^2 = (x-3-2)(x-3+2) = (x-5)(x-1).$$

This method could be described as direct, and always leads to the factored form even where the roots are complex numbers. However, this method does not generalise in the way that expanding out brackets generalises to a much wider range of situations. In particular, it does not generalise to higher order polynomials, or to polynomials in many variables.

Finding the factored form as an intermediate step in solving polynomial equations has become established as the primary contemporary method. This has not always been the case (see Heller, 1940). Indeed, past generations solved equations by seeking direct methods in different cases, e.g. the method of completing the square both solves a quadratic equation and leads to the quadratic formula without the need for factoring. In the past, some students would have been taught to solve cubic equations using the formula, not by guessing a single root and then factoring. The methods taught to students are historically and culturally situated and alternatives exist.

Despite these important differences, a student with an understanding of the relative difficulties of these reversible processes might be tempted to undertake the direct process to verify whether the options for a MC item match the question stem. Such a student might not actually perform the inverse direction, regardless of which method they have been taught. Return again to the example MC item given in the previous section. The reversibility of the factor/expand process can be exploited by testwise students, perhaps as follows. The coefficient of m^3 in the original expression, i.e. 64, can arise only as the product of the first term from each bracket. This immediately eliminates option (1) which does not have a term with m^3, and option (3) where the coefficient is wrong. In this MC item the same reasoning with the constant term (-125) does not eliminate further options. The coefficient of m^2 equals zero in the item stem. Expand option (2) to get the coefficient of m^2 as $-5 \times 16 - 4 \times 20 \neq 0$, so option (2) is eliminated. In the absence of a MC option "*none of the other options*" it is not even necessary to expand out fully to verify that the answer to the factorisation problem is option (4).

Our instrument was administered in an authentic teaching setting, and so we only included two processes: expand/factor and verify/solve, both in limited extent. We only asked students to solve equations of two types, (a) linear equations in a single variable, (b) exponential equations in which the student needed to take logarithms on both sides to reduce the problem to one of class (a). Therefore, the full potential complexity of "solve" was not tested by our study. The full instrument is described below.

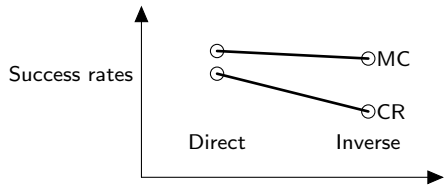


Fig. 1 The expected success rates by format (MC/CR) and direction (direct/inverse)

1.3 Research focus

> When asked to undertake a reversible mathematical processes in multiple choice format do students appear to favour the direct process in both directions?

The specific hypothesis we set out to test was that when faced with a task involving the inverse direction of a reversible mathematical process, students solve a multiple choice (MC) version by verifying the answers by the direct method, not by undertaking the actual inverse calculation. Therefore we expected an asymmetry in the achievement outcomes by item format and process direction. Our study was designed to find out whether there is an item format (MC/CR) and process direction (direct/inverse) interaction for data on students' attempts at undertaking reversible mathematical processes.

Evidence in support of this hypothesis would be the pattern of data shown in Figure 1. Students would be expected to perform better on MC than CR due to the opportunity to select a random response, i.e. guess. It is possible to guess in CR situations but the success rates are likely to be much lower. That said, once we take account of guessing, we might expect students' performance to be about the same on direct MC and CR items. Students would be expected to perform about the same on MC, regardless of direct or inverse process direction. This is because the direct method is available for both and is potentially combined with elimination. (A slightly lower performance might be expected on inverse items as this involves applying direct processes to up to four answers rather than just the question stem.)

Students would be expected to perform significantly worse on inverse CR items compared to inverse MC items. This is because in a CR item there is no longer a direct option. They have to actually perform the (more difficult) inverse task. Therefore, the analysis sought to find a relationship between the percentages of students' correct answers on the multiple-choice format and correct answers on the equivalent constructed responses items.

2 Method

2.1 Participants

129 students enrolled on a foundation programme at a United Kingdom university were invited to participate in the research. The foundation programme helps a variety of students who want to study a science or engineering degree but who have taken an unconventional route through education and find themselves without the subject-specific requirements at the appropriate grades. Consequently, while the cohort did contain some students who has achieved highly in mathematics, it did not contain the normal proportion of high achievers that is typical at university level. However, all students had achieved some success in school mathematics in order to be admitted on to the programme, and most would be expected to go on to attend and complete bachelor degree courses.

Participation in the online test was a compulsory component of the course and contributed to students' final grades, but inclusion in the study was optional. Thirteen students opted out of the study, and a further one student, who opted in, attempted only three items and was omitted from the analysis. This left a total of 116 participants who are included in this report. Although gender differences have been noted in previous research into the effect of question format, our sample included too few women ($N = 26$) for meaningful analysis and this issue is not addressed further.

2.2 Instrument

The instrument was a specially-designed online test suitable for the cohort of foundation programme students. The online test comprised 47 MC and CR items, of which 40 items were included in the analysis. (The additional seven items covered topics that were part of the foundation programme but not relevant to the present study.) We use the term *instrument* to refer to this subset of 40 items in the remainder of the article. The instrument included just two processes from Table 1 as these were the only ones appropriate for this group at the time the study was conducted: expansion/factorisation of simple quadratic/cubic expressions over the integers, and the evaluation of expressions/solving equations in simple cases. The items involved only reversible mathematical process without the problem solving aspects of classical algebra story problems.

For both reversible processes we included items testing the two possible directions in both MC and CR formats. Therefore for every MC item there was an equivalent CR item. The number of items of each type is summarised in Table 2, and the full list of items is in the Appendix.

Writing effective MC items is a non-trivial task, one reason being that all the listed potential answers should be plausible (Friedman et al., 1996). As such, we started with items from `http://mathquest.carroll.edu/`, a publicly available collection of tried and tested items. Each MC item had four options together with the response "none of the other options". For one item "none of the other options" was the correct response.

	Experimental items		Other
Process	Direct	Inverse	
Expand/factor	5	5	
Evaluate/solve	4	6	
Other			4 MC, 3 CR.

Table 2 The number of items included in the instrument by process and direction (direction/inverse) for each format (MC/CR). A total of 40 items was included in the analysis.

MC items were converted into CR items by deleting the response options to create paired versions of the items. Two versions are considered equivalent if and only if the worked solution, written at a level appropriate for the intended student, is invariant. Conversely, different cases in the worked solution requiring different steps indicate the two versions are not equivalent to that student. The precise expressions within steps must vary, of course, but the purpose of the step and the level of detail does not. For example, both $x^2 - 5x + 6 = 0$ and $x^2 - 8x + 7 = 0$ can be solved by factoring, and involve only small integers. The task to solve these two equations would be considered equivalent. The quadratic $x^2 - 6x + 7 = 0$ looks, superficially at least, very similar. While it also has two real solutions it does not factor over the rational numbers, and so a different method of solving it would be needed. In a context in which solving by factoring is the default method, $x^2 - 5x + 6 = 0$ and $x^2 - 6x + 7 = 0$ are not considered equivalent problems. The number of decimal digits in an integer was taken as a proxy for the difficulty of numerical calculations. Numbers of the same order of magnitude were used in corresponding CR and MC items.

In some cases minor changes in the wording of the item were necessary to make explicit what the item was asking. For example, a MC item asking

$$\text{What does } (5x^4)^2 \text{ equal?}$$

could have many correct answers, including $(5x^4)^2$. The MC version does not suffer from this problem as only one of the answers is equivalent to the given expression. Others, such as $25x^6$, arise from a particular mistake. In this case the CR version of the item was as follows.

$$\text{Write } (5x^4)^2 \text{ in the form } ax^n.$$

Where necessary, MC items were similarly reworded for consistency. All the items are in the Appendix.

2.3 Administration

To recruit students to the study the first author attended a lecture and made a short announcement explaining that we would like their permission to use results from a forthcoming test as part of a study to improve the quality of assessment resources in mathematics. Students were informed that

> The online test is a compulsory part of your course. Whether your results are included in the data analysis is up to you. Your decision about this has no impact on your grade for this module, or what you are being asked to do. All data will be completely anonymised prior to analysis. We may link results to background data such as gender and qualifications to help us better understand how to design better online tests.

The online test was administered using the Moodle virtual learning environment as is standard practice at the university. The MC items were implemented directly in Moodle's quiz facility. The CR items were implemented using the STACK system which uses computer algebra to support the assessment process (Sangwin, 2013). The STACK system accepts answers from students in the form of an inputted mathematical expression and then establishes objective mathematical properties of the expression. To do this, tests establish that the student's answer is (i) algebraically *equivalent* to the correct answer and (ii) in the appropriate *form*, (e.g. factored). These are independent objective properties and typically a range of different syntactic expressions satisfy both and hence are considered correct. For further details on the STACK system see Sangwin and Ramsden (2007) and Sangwin (2013).

The participants were familiar with STACK from previous practice assignments. In order to obtain access to the online test, which was compulsory, students had to opt in to or opt out of having their results included in the study. Gender and mathematics achievement data were already available and were matched to students before identifying information was removed, thereby creating an anonymised dataset for the analysis.

Students could sit the online test at any time over the duration of a week, and once logged on were allocated a total of 90 minutes to complete it. (One student was allocated 180 minutes and three students were allocated 112.5 minutes due to specific individual requirements.) The items were presented to students in random order, and students could move between items at will during the test. The STACK system is able to create random versions of a particular item, however this facility was only used on one of the direct items included in the instrument and two of the inverse items. No feedback regarding correctness was available during the test, but students' typed CR expressions were confirmed immediately to them as syntactically valid or invalid. Typed CR expressions were displayed in traditional two dimensional notation and could be modified at any time during the test, e.g. to correct invalidity.

Each question was scored 1 if correct and 0 if incorrect and the results for the 40 items in the instrument were then converted to a percentage for each respondent.

3 Analysis and results

There were three parts to the data analysis. First was ensuring that syntax difficulties had not resulted in unfair automated marking of students' CR responses. Second, reliability and validity checks were undertaken to ensure that the instrument performed as expected. Finally, hypothesis testing was undertaken to explore the differences in accuracy between direct and inverse items in both formats.

3.1 Manual checking of CR input

To ensure the syntax of entering answers did not skew the results we reviewed expressions typed in by students for each of the CR items. Typing in polynomials was unproblematic, although students routinely omitted the star symbol `*` for multiplication. E.g. students type `64m^3-125` rather than `64*m^3-125`. We chose to accept expressions with missing stars. Floating point numbers were rejected immediately as invalid (i.e. not wrong) with very specific feedback, giving students the opportunity to enter an exact answer, e.g. a rational number or surd, instead. Case sensitivity was a problem in some responses: responses in which variables had been entered in the wrong case were marked as wrong.

By reviewing responses for each of the CR items after administration of the instrument we were able to check for any unanticipated responses and decide how these should be scored. Although the criteria need to be specified in advance, criteria can be changed and the students' answers reassessed at a later time. This procedure corresponds to reviewing MC options to see if a particular item is functioning well in a test.

3.2 Reliability and validity

The coefficient of internal consistency (Cronbach's alpha) was high, $\alpha = .91$, suggesting the instrument performed reliably. We also considered the internal consistency of subsets of items. For the direct MC items ($N = 9$) the coefficient was $\alpha = .71$, for the inverse MC items ($N = 11$) it was $\alpha = .69$, for the direct CR items ($N = 9$) it was $\alpha = .80$, and for the inverse CR items ($N = 11$) it was $\alpha = .77$. The internal consistency coefficients are lower for the subsets than for all items taken together, which is to be expected given that the value of Cronbach's α is dependent on the number of items in a test, but nonetheless provide support for the performance of the instrument.

We also investigated the consistency of the items in terms of the CR and MC formats. Above we noted Kamps and van Lint's (1975) reported correlation coefficient between CR and MC formats of $r = .57$. This coefficient does not demonstrate a lack of notable item effect as 68% of the variance is left unexplained. In the present study we obtained a much higher correlation coefficient, $r = .84$. This accounts for 71% of the variance and provides reasonable support that overall format effect was not large.

Exploratory factor analysis resulted in all items loading on a single component, supporting the unidimensionality of the instrument. As such, a composite score was calculated for each student across the 40 dichotomous items included in the study, which was expressed as a percentage. The mean overall score was 68.8% with a standard deviation of 19.1%.

To investigate criterion validity, the students' composite scores were correlated with their scores for other assessments administered on the module. These were a second computer-based test on the topic of differentiation, a paper-based test sampling a wide range of mathematical topics from across the module, and a

	Online test 2	Paper test	Exam
Composite score	.58	.58	.59
Online test 2		.50	.53
Paper test			.81

Table 3 Correlation matrix of Pearson product-moment coefficients between student scores across four module assessments. *Composite score* is the mean score across the 40 items used in the study.

Format	Direction	Mean	Sd
CR	Direct	69.1	24.2
CR	Inverse	60.4	23.0
MC	Direct	77.5	22.1
MC	Inverse	73.8	19.1

Table 4 Success data as percentage achievement by format and direction.

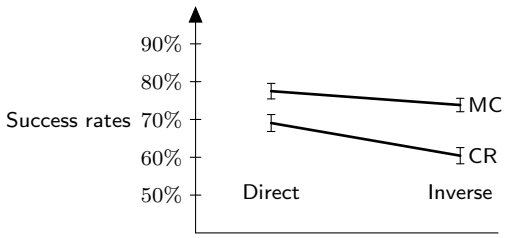


Fig. 2 Success rates. Error bars represent ±1 SE of the mean.

synoptic examination also sampling a wide range of topics. Complete results across the three assessments were available for 110 of the students who participated in the research. The correlation matrix is shown in Table 3. The highest correlation coefficient, $r = .81$, is between a paper-based test and synoptic exam, both of which sampled widely across mathematical topics. The correlations between the computer-based tests and other assessments are lower, ranging from $r = .50$ to $r = .59$, which is to be expected because each focussed on specific topics (algebraic manipulation and differentiation respectively). Nevertheless, taken together these correlation coefficients support the overall validity of the composite scores as a measure of students' mathematical achievement.

3.3 Effect of direction and format

The mean scores by item format and direction are summarised in Table 4 and Figure 2. Mean scores were higher for MC than CR items overall in line with our prediction. In addition, for each format, mean scores were higher for direct than inverse items which is also in line with our prediction. To test whether these differences were significant the data were subjected to a 2 (format: MC, CR) by 2 (direction: direct, inverse) Analysis of Variance (ANOVA), where both factors

	CR Direct	MC Inverse	CR Inverse
MC Direct	**.744**	*.600*	.595
CR Direct		.636	*.673*
MC Inverse			**.786**

Table 5 Correlation matrix of Pearson product-moment coefficients between student scores across the four question types (format × direction). Within-direction coefficients are shown in bold, within-format coefficients are shown in italics.

were within subjects. As expected, this revealed a main effect of format, with MC items being answered significantly more accurately (75.7%) than CR items (64.7%), $F(1, 115) = 102.371$, $p < .001$, $\eta_p^2 = .471$.

The difference in overall success rates between CR and MC of 11% is somewhat smaller than the 20%, which might be attributed to pure guessing between 5 equally likely options. However, the higher overall success rates reduce the potential effect of guessing. To better estimate the effect of guessing, assume a student has a 65% chance of knowing how to complete an item correctly. This is a realistic scenario given the CR data. We assume that in 35% of cases a student will not know how to proceed, and will therefore guess, with a 1/5=20% chance of success, giving an overall guessing advantage of $35\% \times 20\% \approx 7\%$ for MC over CR. If a student ignores the "none of the others" option and strategically eliminates one further option (or eliminates two options), then guesses from the remaining three, their expected overall guessing advantage would be $35\% \times 33\% \approx 12\%$. We therefore consider the difference of 11% between CR and MC to be consistent with partial guessing in cases where students do not otherwise know how to solve a CR question.

There was also a significant format by direction interaction, $F(1, 115) = 6.892$, $p = .010$, $\eta_p^2 = .057$. This interaction was investigated with a series of planned comparisons. For CR items, accuracy was significantly higher on direct compared to inverse items, 69.1% versus 60.4%, $t(115) = 4.861$, $p < .001$, $d = 0.451$. A smaller, but still significant, effect was also observed for MC items, 77.5% versus 73.8%, $t(115) = 2.125$, $p = .036$, $d = 0.197$. To investigate whether the effect of direction was significantly different across the two formats we calculated the differences between scores for MC and CR by direction. For direct items the mean difference was 8.4% and for inverse items the mean difference was 13.5%, and this difference was significant, $t(115) = -2.666, p = .009$. Therefore students' relative performance across direct and inverse items was significantly different across the two formats, with the relative performance lower for CR items.

A consideration of the correlation matrix of scores for the four types of question (format × direction) provides further insight, as shown in Table 5. The correlations within direction (shown in bold in Table 5) are stronger than the correlations within format (shown in italics in Table 5). This is consistent with the hypothesis that the items are equivalent across formats, and that direction is driving the differences in achievement across the question types.

We also considered the performance of students on direct and inverse items across the two formats at the individual level. For direct items, 54.3% of participants scored more highly on MC than CR items, 32.0% scored the same across both

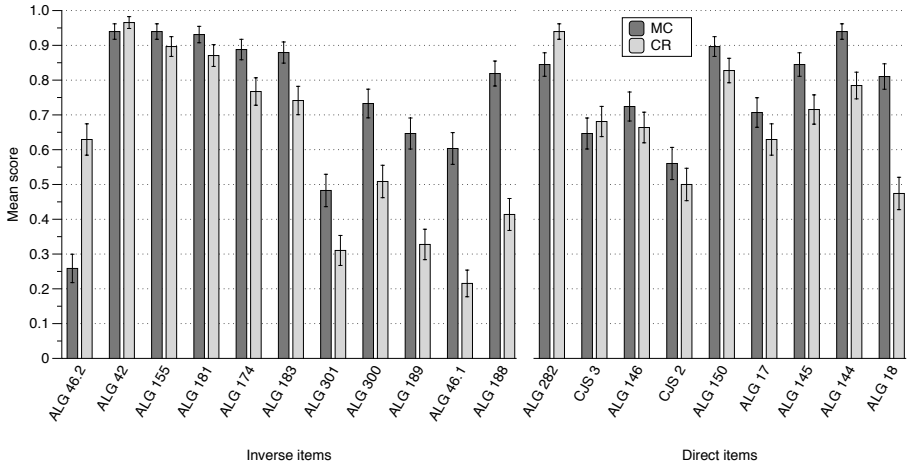


Fig. 3 Performance of individual items by format. Items are shown in left-to-right order of the difference in means scores between CR and MC formats.

formats, and only 13.8% scored more highly on CR than MC items. For inverse items the figures were 74.1%, 16.4% and 9.5% respectively.

We also considered the items in our instrument individually, as shown in Figure 3. Only one question, ALG 46.2, performed anomalously compared to the overall trend. This question asks students *"What is the solution set:* $2(x-3) = 5x - 3(x+2)$*?"*. Gathering like terms gives $0 = 0$ indicating that any value of x satisfies the equation. In the MC condition students were given the choice between three sets each containing one specified real number, the option *"{ all real numbers }"* and *"No solutions"*. In the CR condition students were expected to give a set of numbers representing the solutions. Students were also instructed to *"Type in* $\{R\}$ *if there is more than one solution, and* $\{\}$ *if there are no solutions."*. Only 27,5% of students answered this question correctly in MC format, but 71.6% of students answered this correctly as a CR question. Without this question our trend showing an asymmetry of achievement would be more pronounced.

These analyses support our hypothesis of an asymmetry of achievement. The mechanism we propose for this asymmetry is that students carry out direct processes on the provided answers to inverse MC items.

3.4 Role of mathematical achievement

Finally, we undertook an unplanned analysis to explore whether students' overall performance on the instrument interacted with their performance on direct and inverse items across the two formats. Recall that the mean overall score was 68.8% with a standard deviation of 19.1%. Students who scored below the mean ($N = 51$) were assigned to a low-achieving group and those who scored above the mean ($N = 65$) were assigned to a high-achieving group. As for the main analysis, we calculated differences between scores for MC and CR by direction for each group.

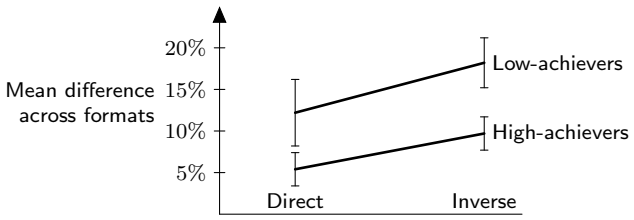


Fig. 4 Mean differences across formats for low- and high-achievers. Error bars represent ±1 SE of the mean.

For the low-achieving and high-achieving groups the mean difference of direct items between formats was 12.2% and 5.4% respectively; the mean difference of inverse items between formats was 18.2% and 9.7% respectively, as shown in Figure 4.

To investigate group differences we conducted a mixed between-within subjects ANOVA, with mean difference between MC and CR scores for each direction (direct, inverse) as the within-subjects factor, and achievement (low, high) as the between-subjects factor. There was a significant main effect for mean differences between MC and CR by direction $F(1, 114) = 7.243, p = .008, \eta_p^2 = .060$. However the interaction was not significant, $F(1, 114) = .199.p = .657$, $\eta_p^2 = .002$, suggesting the main effect was due to better performance of both groups on direct over inverse, as shown in Figure 4. Therefore there was no evidence to support a difference between the low- and high-achievers in terms of performance on direct and inverse items across both formats.

4 Discussion

We compared students' performance on MC and CR items used in an online test as part of a compulsory summative assessment. We found that, overall, students performed better on MC than on CR items. Some of this difference is likely to be accounted for by the use of guessing in MC items. Critically, however, the improved performance for MC items was greater for items intended to test competence with inverse processes compared to items intended to test direct processes. Finding this asymmetry supports the hypothesis that when faced with an item involving the inverse direction of a reversible mathematical process, students commonly solve a MC version by verifying the options using a direct method, and not by undertaking the actual inverse calculation. Moreover, this finding is robust across low- and high-achievers: item format and direction did not appear to affect these two groups of learners differently. These results present a serious challenge to the use of MC items for assessing reversible mathematical processes because it cannot be determined by the item writer exactly what is being assessed. The study reported here focussed on the processes of expansion/factorisation and evaluation/solving but the principle can be extended to other processes such as those listed in Table 1.

It is likely in practice that students will often, quite understandably and rationally, take the easiest path when faced with a MC item that involves reversible

processes. This has worrying implications not just for valid assessment of students' knowledge and skills, but for the impact such assessment has on their learning and future mathematical development. It is likely that a student who has succeeded only on MC items at an earlier stage via direct verification would be at a serious disadvantage when confronted with a CR item at a later date. On this hypothesis, a mathematics educator who relies on the MC format for assessing reversible processes may be performing a serious disservice to his or her students in the longer term. Designers of online materials which rely on MC face the same dilemma between the technical simplicity of MC and the educational validity of CR.

It is therefore recommended that MC formats be avoided for the assessment of reversible mathematical processes. One option is to use CR formats, which are relatively easy to score reliably (Newton, 1996), or can be implemented online and scored automatically (Sangwin, 2013), as was the case here.

4.1 Limitations

While the study yielded a clear and predicted result, caution must be exercised when interpreting the generality of the finding. We highlight three main limitations.

First, our findings apply exclusively and explicitly to reversible processes only. MC items are not generally invalid for assessing mathematics, and both the authors use them in their teaching and assessing of mathematics at university level. Indeed, there are many contexts in which well designed MC items are more appropriate than other question formats. For example, the popular Calculus Conceptual Inventory (Epstein, 2013) has appealing face validity and that would be lost if converted into a CR format. This is because items in the Calculus Conceptual Inventory tend to avoid calculation, and therefore the issue of reversible processes, focussing instead on underlying principles.

Second, the study used a modest sample of students ($N = 116$) from a single cohort at a single university. We are confident, due to the theoretical reasons stated earlier, as well as discovering that our main result was robust across low- and high-achievers, that the same methods applied to different cohorts in different universities would lead to the same broad finding. Nevertheless, we cannot claim that our sample is representative of the broader population of students undertaking mathematics modules at universities around the world. In particular, the cohort was taking a foundational course as a prerequisite for embarking on bachelor degrees, mainly in engineering and the sciences. Therefore we would expect a broader variation and lower mean achievement in mathematics than for other samples of undergraduates. Moreover, most of the participants were male (78%), which may have slightly inflated the MC scores (Hassmén & Hunt, 1994; Livingston & Rupp, 2004; Mazzeo et al., 1993), and this also barred us from investigating hypothesised gender effects.

Third, CR items that require an online response, such as the STACK system used here, raise the difficulty of students needing to learn specialised syntax to enter their answers. Further discussion of this issue can be found in Sangwin and Ramsden (2007). We reported that students' performance was worse on the CR

than the MC items, and this was despite our manual checking of student responses as described in the methods section above. Without such checking the disparity is likely to have been greater still. The extent to which syntax gets in the way of students providing mathematical answers presents a validity threat to online CR assessment systems. We chose to use an online system to gather the data for our study because our students were already using this system, making it an authentic assessment experience. In addition, it offers an efficient, reliable and convenient way to gather the data from a large cohort of students. However, we cannot be certain whether the effect is confounded with the use of technology as opposed to working on paper.

These limitations are readily overcome in future work. The same methods can be applied to different reversible processes, using different samples of the student population, and implementing the CR items using different online systems or pencil and paper. (However we acknowledge that since online assessment is becoming more common conducting this study using pencil and paper may limit future relevance.)

5 Conclusion

Our research found evidence for an item format (MC/CR) and process direction (direct/inverse) interaction for reversible mathematical processes. This evidence supports the hypothesis that when faced with a task involving the inverse direction of a reversible mathematical process, students solve a multiple choice (MC) version by verifying the answers by the direct method, not by undertaking the actual inverse calculation. It might be that MC items could provide an advantage to lower achievers in particular, however we found no evidence to support this hypothesis.

Should mathematics be assessed using MC items? If the focus of assessment is on reversible processes then the answer is no. Presented with this format students will take the easiest path, performing inverse processes on answers rather than a direct process on the item stem. Such a strategy allows examinees to perform above chance by side-stepping what the item writer intends to assess. Instead, reversible mathematical processes should be assessed using CR or other open-ended item formats.

Acknowledgements Thanks to Prof. K. Cline for granting permission to use the Mathquest materials as a starting point for our instrument. Thanks to Dr Matthew Inglis for advice on statistical analysis. Thanks to the reviewers for their helpful comments.

References

Bridgeman B (1992) A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement 29(3):253–271

Brousseau G (1997) Theory of Didactical Situations in Mathematics: didactiques des mathématiques, 1970–1990. Kluwer, n. Balacheff, M. Cooper, R. Sutherland, and V. Warfield (Trans.)

Davenport JH, Siret Y, Tournier E (1993) Computer algebra: systems and algorithms for algebraic computation. Academic Press Professional

Epstein JM (2013) The calculus concept inventory—measurement of the effect of teaching methodology in mathematics. Notices of the American Mathematical Society 60(8):1018–1026

Friedman DE, Bennett RE, Katz IR, Berger AE (1996) Differences in strategies used to solve stem-equivalent constructed-response and multiple-choice sat-mathematics items. Research Report 96-20, The Educational Testing Service

Goodwin KS, Ostrom L, Scott KW (2009) Gender differences in mathematics self-efficacy and back substitution in multiple-choice assessment. Journal of Adult Education 38(1):22–42

Hassmén P, Hunt DP (1994) Human self-assessment in multiple choice. Journal of Educational Measurement 31(2):149–160

Heller HF (1940) Concerning the evolution of the topic of factoring in textbooks of elementary algebra published in England and the United States from 1631 to 1890. Phd, Columbia University

Iannone P, Simpson A (2012) Mapping University Mathematics Assessment Practices. University of East Anglia

Kamps HJL, van Lint JH (1975) A comparison of a classical calculus test with a similar multiple choice test. Educational Studies in Mathematics 6(3):259–271

Livingston SA, Rupp SL (2004) Performance of men and women on multiple-choice and constructed-response tests for beginning teachers. Research Report 04-48, Educational Testing Services

Martinez ME (1999) Cognition and the question of test item format. Educational Psychologist 34(4):207–218, DOI 10.1207/s15326985ep3404$_2$

Mazzeo J, Schmitt AP, Bleistein CA (1993) Sex-related performance differences on constructed-response and multiple-choice selections of advanced placement examinations. College Board Report 92-7, College Entrance Examination Board, New York

Risch RH (1969) The problem of integration in finite terms. Transactions of the American Mathematical Society 139:167–189

Rittle-Johnson B, Siegler R, Alibali M (2001) Developing conceptual understanding and procedural skill in mathematics: An iterative process. Journal of Educational Psychology 93:346–362

Sangwin CJ (2013) Computer Aided Assessment of Mathematics. Oxford University Press

Sangwin CJ, Ramsden P (2007) Linear syntax for communicating elementary mathematics. Journal of Symbolic Computation 42(9):902–934, DOI 10.1016/j.jsc.2007.07.002

Shepard LS (2008) Commentary on the national mathematics advisory panel recommendations on assessment. Educational Researcher 37(9):602–609, DOI 10.3102/0013189X08328001

References

Berube, C. T. (2004). Are standards preventing good teaching? *Clearing House, 77*, 264–267.

Bisson, M., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education, 2*, 141–164.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253–271.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies, 52*, 29–46.

Brousseau, G. (1997). *Theory of Didactical Situations in Mathematics: Didactiques des Mathématiques, 1970–1990.* London: Kluwer Academic Publishers.

Davenport, J. H., Siret, Y., & Tournier, E. (1993). *Computer Algebra: Systems and Algorithms for Algebraic Computation.* London: Academic Press Professional.

Epstein, J. M. (2013). The calculus concept inventory—measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society, 60*, 1018–1026.

Friedman, D. E., Bennett, R. E., Katz, I. R., & Berger, A. E. (1996). *Differences in Strategies Used to Solve Stem-Equivalent Constructed-Response and Multiple-Choice SAT-mathematics Items. Research Report 96–20.* New York: College Entrance Examination Board.

Goodwin, K. S., Ostrom, L., & Scott, K. W. (2009). Gender differences in mathematics self-efficacy and back substitution in multiple-choice assessment. *Journal of Adult Education, 38*, 22–42.

Hassmén, P. & Hunt, D. P. (1994). Human self-assessment in multiple choice. *Journal of Educational Measurement, 31*, 149–160.

Heller, H. F. (1940). *Concerning the Evolution of the Topic of Factoring in Textbooks of Elementary Algebra Published in England and the United States from 1631 to 1890.* PhD thesis: Columbia University.

Iannone, P. & Simpson, A. (2012). *Mapping University Mathematics Assessment Practices.* Norwich, UK: University of East Anglia.

Kamps, H. J. L. & van Lint, J. H. (1975). A comparison of a classical calculus test with a similar multiple choice test. *Educational Studies in Mathematics, 6*, 259–271.

Livingston, S. A. & Rupp, S. L. (2004). *Performance of Men and Women on Multiple-Choice and Constructed-Response Tests for Beginning Teachers. Research Report 04-48.* Princeton, NJ: Educational Testing Services.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207–218.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related Performance Differences on Constructed-Response and Multiple-Choice Selections of Advanced Placement Examinations. College Board Report 92-7.* New York: College Entrance Examination Board.

Newton, P. & Shaw, S. (2014). *Validity in Educational and Psychological Assessment.* London: SAGE.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal, 22*, 405–420.

Noyes, A., Wake, G., Drake, P., & Murphy, R. (2011). *Evaluating Mathematics Pathways: Final Report. DfE Research Report 143.* London: Department for Education.

Risch, R. H. (1969). The problem of integration in finite terms. *Transactions of the American Mathematical Society, 139*, 167–189.

Rittle-Johnson, B. & Siegler, R. (1999). The relation between conceptual and procedural knowledge in learning mathematics: a review. *Journal of Educational Psychology, 91*, 175-189.

Rittle-Johnson, B., Siegler, R., & Alibali, M. (2001). Developing conceptual understanding and procedural skill in mathematics: an iterative process. *Journal of Educational Psychology, 93*, 346–362.

Sangwin, C. J. (2013). *Computer Aided Assessment of Mathematics.* Oxford: Oxford University Press.

Sangwin, C. J. & Ramsden, P. (2007). Linear syntax for communicating elementary mathematics. *Journal of Symbolic Computation, 42*, 902–934.

Shepard, L. S. (2008). Commentary on the national mathematics advisory panel recommendations on assessment. *Educational Researcher, 37*, 602–609.

Spencer, S., Steele, C., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4–28.

Swan, M. & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer, 2*(5), 1–41.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3–13: International Journal of Primary, Elementary and Early Years Education, 29*, 17–21.

A Items used in the instrument

Items used in the test are shown below. In the case of expand/factor and solve/evaluate equivalent versions of each item were used for both CR and MC items. We have used items, with permission, from `http://mathquest.carroll.edu/`, and the tags such as ALG 144 indicate which items we have taken. Responses have been omitted here, but are available in the document online. The tags CJS and IJ indicate the author of additional items where this catalogue did not contain sufficient, particularly in the evaluate class.

Expand/factor

ALG 144 (*direct*): Expand and simplify: $(2x+5)(3x+2)$.

ALG 145 (*direct*): Expand and simplify: $(7x+2)(x^2+8x-3)$.

ALG 146 (*direct*): Expand and simplify $(3x-5)^2$.

ALG 150 (*direct*): Expand and simplify $(3x-4)(3x+4)$.

ALG 282 (*direct*): Multiply and simplify: $(5+\sqrt{3})(5-\sqrt{3})$.

ALG 174 (*inverse*): What is the greatest common factor of the terms of $20x^2+28x$?

ALG 181 (*inverse*): Factor $x^2-7x+12$.

ALG 183 (*inverse*): Factor $x^2+10x-11$.

ALG 188 (*inverse*): Factorise: $45m^2-20$.

ALG 189 (*inverse*): Factorise: $64m^3-125$.

Evaluate/solve

ALG 17 (*direct*): Evaluate: $3x^2 - 7xy + 4y^2$ when $x = -2$ and $y = 3$.

ALG 18 (*direct*): Evaluate: $\frac{5ab^2}{2a^2-3b}$ when $a = 5$ and $b = -1$.

CJS 2 (*direct*): Evaluate $\frac{\log_{10}(x^{400})}{70}$ when $x = 10^{70}$.

CJS 3 (*direct*): Substitute $x = -1$, $y = 2$ and $z = -3$ into

$$\frac{y-z}{yz} + \frac{z-x}{zx} + \frac{x-y}{xy}$$

and calculate the result.

ALG 155 (*inverse*): What is x if $2x + 5 = 0$?

ALG 42 (*inverse*): Solve for x: $x + 7 = 8$.

ALG 46(1) (*inverse*): What is the solution set: $2(x - 3) = 4x - 3(x + 2)$?

ALG 46(2) (*inverse*): What is the solution set: $2(x - 3) = 5x - 3(x + 2)$?

ALG 300 (*inverse*): Solve: $2^x = 5$.

ALG 301 (*inverse*): Solve $3^{x-2} = 5^{4x}$.

Other

These questions were included in the test but not in the analysis for the study.

ALG 117: Which of the following is equivalent to

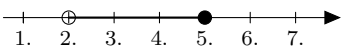
$$\frac{x^{-2}y^3z^{-4}}{x^{-3}y^5z^5}$$

when written in the form $x^a y^b z^c$?

ALG 160: If $x = -\frac{5}{10}$ is a zero, then the corresponding factor is: ...

ALG 192: If $(x - 2)(x + 1) = 10$ then find x.
Note: this item was not included in the study as a correct solution requires both expand and factor.

ALG 4: Which inequality corresponds to this graph? [MCQ choices given]



CJS 1: A university has 6 times as many students as professors. If S represents the number of students and P represents the number of professors, which of the following equations expresses the relationship between S and P?

IJ 1: Suppose that $\ln(2) = a$ and $\ln(5) = b$. How might $\ln(10)$ be written?

IJ 2: Express $(\log(x) - \log(y)) + 3\log(z)$ as a single logarithm.

Comparison of MC and CR questions

As examples of our question rewriting, the following is the MC version of ALG 188

> Factor: $45m^2 - 20$.
> (a) $(7m - 5)(7m + 5)$
> (b) $5(9m - 4)(9m + 4)$
> (c) $5(3m - 4)(3m + 4)$
> (d) $5(3m - 2)(3m + 2)$

The CR response version is

> Factor: $45m^2 - 20$.

This is typical: to create the CR versions we took an existing MC question and replaced the options with an answer box into which the student is expected to type their answer.