Teaching Mathematics and its Applications: An International Journal of the IMA (2022) **00**, 1–12
https://doi.org/10.1093/teamat/hrac002 Advance Access publication

Design principles for final answer assessment in linear algebra: implications for digital testing

Alisa J. Veale[†,*], Tracy S. Craig[‡]

[†] *Centre of Expertise in Learning and Teaching, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

[‡] *Department of Applied Mathematics, University of Twente Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

[*] *Corresponding author. Email:* alisa.veale@utwente.nl

[Received October 2021; accepted January 2022]

Digital testing, such as multiple-choice questions and final answer items, offers many advantages in higher educational assessment practices. Well-designed digital grading is more reliable and faster than hand grading and is scalable to larger classes. The validity of digital grading is open to criticism, particularly in mathematics where much of mathematics is based on processes and reasoning and not on the final answer achieved. At a technical university in the Netherlands, we have been increasing our use of digital short answer testing in calculus and linear algebra for service mathematics. To assess the validity of this mode of assessment, we graded a linear algebra test in two ways: short answer grading (where answers were considered either correct or incorrect) and the so-called 'hypothetical grading', where we assigned a grade based on the fully worked solution. Certain types of items proved to be more suitable for short answer (and hence digital) testing than others. We concluded our analysis with a set of design principles for digital or short answer testing in linear algebra.

1. Introduction

Digital testing of mathematics holds many advantages over pencil and paper grading with its speed, opportunity for immediate automated feedback, potential for scaling up and not being susceptible to grader error (Aarts, 2018; Sangwin, 2019a). Use of digital testing, or computer-aided assessment (Lawson, 2002; Broughton *et al.*, 2013), is on the rise as class sizes increase globally and the tools and platforms available becoming increasingly sophisticated (Koomen & Zoanetti, 2018; Sangwin & Köcher, 2016).

Despite the many advantages, there are concerns surrounding the digital testing of mathematics (Lawson, 2002; Robinson *et al.*, 2012; McGuire *et al.*, 2002). These concerns are frequently grounded in the importance of being able to assess process in any mathematical task rather than simply the production of an answer. Current digital testing, however, can generally only assess the final answer to the question, rather than the process, with some exceptions such as 'reasoning by equivalence' proof checking

© The Author(s) 2022. Published by Oxford University Press on behalf of The Institute of Mathematics and its Applications. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

(Sangwin, 2019b), embedded answer items (Martins, 2018) and freehand sketching (Yerushalmy *et al.*, 2017). Designing good assessment items for final answer or digital grading is more challenging than designing items for traditional partial credit grading. As such, a set of principles informing the design of valid and reliable test items which successfully assess learning outcomes in the absence of visible student mathematical process are valuable.

At the University of Twente, digital testing has played a role in certain of the service mathematics courses since the academic year of 2015/2016, specifically one calculus course and a linear algebra course. As Koomen & Zoanetti (2018) point out, for a paper and pencil partial-credit test one needs only a small team of experts, a word processor and a printer. A shift to technology-based testing, particularly on a large scale, requires 'organisational transformation to coordinate the new and emerging divisions of labour that technology generates' (Koomen & Zoanetti, 2018, p. 201). In our case, one such transformation is the acquisition of Chromebooks for students to use identical devices with controlled access to predetermined resources. This is in contrast to the 'bring your own device' model which has its advantages but also significant disadvantages (Moccozet *et al.*, 2018). This acquisition process is ongoing, and currently more students are taking service Mathematics subjects than Chromebooks are available, occasionally causing a paper-proxy to be used for digital testing until infrastructure catches up.

In June 2019, a Linear Algebra test was written, where 50% of the questions were only judged on their final answer. During the 10-week module, students were prepared for these final answer questions, through in-class discussions and Q&A sessions and through written examples of how students can check their final answer for correctness. Two of the four cohorts of students that wrote the test handed in their rough work for analysis purposes to investigate the quality of the exam, especially that of the final answer questions. The analysis discussed below provided insight allowing us to refine the existing design principles for the creation of good quality items for final answer (preferably digital) testing.

2. Research questions

The analysis carried out in this study considered the validity of grading the items by their final answer only by comparing that mode of grading to a partial credit mode in order to determine the quality. The objective in this study was a set of design principles for items for digital grading, specifically final answer items and in the context of linear algebra. To support this objective, we asked two research questions.

1. What do the quantitative differences between grades achieved through final answer grading versus partial credit grading tell us about the quality of the items?

2. What types of errors are observed in students' working and how does that inform item design?

The answers of the two research questions allow us to refine the existing design principles.

3. Prior design principles

Design based research consists of different phases. The model produced by McKenney & Reeves (2018) consists of three main phases: analysis, design and evaluation. This paper represents part of the evaluation phase of the cycle. Having developed a first set of design principles from the linear algebra test of the previous year and research done on Calculus at the University of Twente (Aarts, 2018; Lochner, 2019) we applied and evaluated these principles in this exam. For the reader, we will recall the design principles in this section, reflect on the implementation process in the methodology section and then describe the outcomes in the results and discussions sections.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

The design principles, in no particular order, consisted of the following.

- Limit the number of points assigned to an item to a maximum of three.
- An item should be a one step process not susceptible to careless error OR the answers should be checkable and students should have had the opportunity to learn how to do this.
- The algebra involved should be minor, testing the core concept of the question, rather than algebraic manipulation.
- As the learning goals that can be tested in this way are limited, a maximum of 50% of the grade should be based on final answer questions.

4. Research methodology

Four cohorts of students enrolled in an introductory linear algebra course in March 2019: physics, electrical engineering, advanced technology and mechanical engineering students. The course lasted one academic quarter of 10 weeks. For the purposes of this analysis, the two cohorts of physics (N = 47) and electrical engineering (N = 77) students were chosen to have the grading of their work analysed for this study.

The students' work was first graded for summative purposes. Half of the paper (18 of 36 points) was in the form of fully worked solutions which could receive partial credit. The other half was graded only on the final answer. The final-answer items can be found in the appendix, where Items 1, 3, 4 and 5 were graded on the final answer only and Items 2, 6 and 7 were graded on their fully worked solutions. The university is in the process of increasing the number of devices available for use in a digital test. Due to logistic difficulties, in this instance, all students completed the 'digital' part of the test on paper, writing their answers into blocks on an answer sheet; this process was already familiar to the students. The analysis discussed below therefore was of paper-proxy for digital testing which had the dual result of providing insight into final answer testing as well as the effectiveness of having a paper-proxy.

The final answer questions were re-graded using a grading scheme which allowed partial credit, hereafter termed 'hypothetical grading' or 'partial credit grading'. (For related work, see McGuire *et al.*, 2002; Ashton *et al.*, 2006; Rane & MacKenzie, 2020.) The work being graded this way was necessarily 'rough work' as the students had not written it with an expectation of it being graded. Consequences included absence of rough work for some students, work that was hard to read and incomplete rough work. The 124 students whose work was analysed included only those who submitted complete and readable rough work. We were interested not only in whether the grades differed but also in why.

5. Results

5.1. Quantitative differences between grades achieved through final answer grading versus partial credit grading

We present the data on the two forms of grading from three viewpoints, that is, difference in total grade (overall difference), item-by-item analysis and the pass-fail distinction.

5.1.1. Overall difference. Table 1 presents statistics on the differences in total grade for the test between final answer grading and (hypothetical) partial credit grading. The maximum grade achievable in the test was 36, with 18 of those points assigned to final answer items. The column 'point difference' provides hypothetical grade—final answer grade. In total, 53% of students have fewer than 2-point difference

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

TABLE 1. *Total grade difference: hypothetical partial credit—final answer grade*

	Per grade point			
Point difference	Number of students	% of students	Cumulative number of students	% Cumulative students
<0	9	7%	9	7%
=0	36	29%	45	36%
0.5	2	2%	47	38%
1 & 1.5	19	15%	66	53%
2 & 2.5	21	17%	87	70%
3 & 3.5	13	10%	100	81%
4 & 4.5	12	10%	112	90%
5 & 5.5	8	6%	120	97%
6	2	2%	122	98%
7.5	2	2%	124	100%

between the two types of grading across the whole test and 70% have fewer than a 3-point difference across the whole test. Moreover, 29% of students did not have a difference in points between the two grading schemes, larger than any other group in the column '% of students'. A total of 7% of the cohort achieved a lower grade in the hypothetical grading mode.

The greatest difference in grading modes was 7.5 points, incurred by two students. The hypothetical grades remained failing grades; we explain this situation in the Discussion below. Four students experienced a two-symbol difference (20% difference) in final grade. A total of 58% of students experienced no change in final grade whatsoever. The items discussed below can be found in the appendix.

5.1.2. Per item. We compared the results of the two forms of grading item by item to determine if particular types of items exhibited greater sensitivity to variation. All the correlations for Questions 1a, 1b, 3, 5a and 5b were above the 0.75 mark, showing a relatively high correlation for two grading schemes that assign points differently. The grades for Questions 4a and 4b were more difficult to correlate as there was a fundamental difference between the two grading schemes, greater than that of the other questions, discussed elsewhere in this article.

5.1.3. Pass versus fail. A difference in points is also important for the boundary level of passing or failing. In total, there were seven students who would have passed the test if they were graded using the hypothetical grading instead of the final answer grading. At the same time, there is one student who would have failed if the hypothetical grading was used instead of the final answer grading.

5.2. What types of errors are observed in students' working

As reported in the method section, errors were recorded and placed either into the category of arithmetic or understanding errors. The few instances of copying errors were bundled with the arithmetic errors in that they were careless mistakes, not errors of conceptual understanding. In total, 42 (34%) of the scripts were checked by a lecturer, not only for the hypothetical grading but also for these classifications of types of errors. The interrater reliability was determined to be 98.5%. As with Research Question 1 we take an 'overall' look at the data and then consider them item by item.

TABLE 2. *Per question difference in grading modes*

	Q1a		Q1b		Q3		Q5a		Q5b		
Diff	Und	Arith	Und	Arith	Und	Arith	Und	Arith	Und	Arith	Tot
-2	—	—	8	—	2	—	—	—	—	—	10
-1	—	—	—	1	—	—	—	—	—	—	1
+0.5	—	—	—	—	—	—	—	—	9	2	11
+1	—	4	2	10	13	—	6	2	2	3	42
+1.5	—	—	—	—	—	—	—	—	4	15	19
+2	—	—	—	1	4	10	9	5	—	1	30
+3	—	—	—	—	—	1	—	3	—	—	4
Tot	0	4	10	12	19	11	15	10	15	21	117

5.2.1. Overall: arithmetic and understanding errors. In Table 2 below, we show numbers of students either losing or gaining points in the shift to hypothetical partial credit grading. We have omitted Questions 4a and 4b from the table for reasons discussed later. Across the remaining questions (1a, 1b, 3, 5a and 5b) a total of 117 grade-influencing errors were recorded, 59 of them errors of conceptual understanding and 58 of them arithmetic or careless in nature.

5.2.2. Per item. For two points, Question 1a asked the student to find the determinant of a three-by-three matrix. The prevalence of zeros in the matrix made the determinant easy to compute as long as the sign of the (2,3)-minor is kept in mind. The item fits with the existing design principles by testing a single step process not readily susceptible to careless errors. It is checkable by computing the determinant a second time, expanding about a different row or column to the first time. Out of 124 students, only 4 made a small arithmetic error, causing them to get one out of the two available points in the hypothetical grading, while they had zero in the digital grading. When comparing the two types of grading, 94 (75.8%) students got full credit in both modes and 26 (21%) students got no points in both modes.

For two points, Question 1b asked the student to evaluate the determinant of the inverse of the matrix in 1a. Many students chose to first invert the matrix and then find the determinant, which was far more prone to error than the procedure of taking the reciprocal of the determinant in Question 1a. With these two methods available to students, they could arguably check their answer, which aligns with our design principles. Final answer grading gave full credit to 1b if the answer were correct or if it were the reciprocal of the answer in 1a.

Students tended to get a lower result in partial credit grading than in final answer grading in this item (Question 1b) more than in any other on the test, due to the many arithmetic errors. One copy error caused the student to gain two points in the hypothetical partial credit grading, as when they put their answer on the final sheet, they forgot to include the minus. When comparing the two types of grading, 78 (62.9%) students got everything correct in both modes and 24 (19.4%) students got everything incorrect in both modes.

For three points, Question 3 asked students to indicate which of the four given vectors were elements of set T, the span of a two-dimensional basis. In other words, students were asked to find out which of the vectors were linear combinations of the vectors in the basis of T. An efficient method for determining which vectors qualify is to augment the basis against all four vectors and reduce to find the consistent systems. A less efficient approach would be to carry out the same process separately for each vector,

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

which is often what students did. The item is aligned with the design principles by testing a single concept. While susceptible to careless error the result is checkable by determining coordinates of the vectors with respect to the basis of T and checking that the linear combination does work.

The final answer grading awarded full credit for identifying the two (and only those two) vectors which are elements of T and no credit for any other answer. The hypothetical grading awarded one point each for setting up the augmented matrix (or four small ones), reducing to echelon form and interpreting the result. For a student who got zero for the final answer grading, partial credit was possible if reducing was done incorrectly or if interpretation of the result was not correct. For a student who got full credit (three points) for the final answer grading it was possible to lose points in the hypothetical grading if the answer was effectively accidentally correct.

In the grading of Question 3, 68 students (54.8%) gained full credits in both modes of grading, 26 students (21%) got zero in both modes of grading. Thirty (24.2%) students scored differently across the two modes of grading; 28 of those scoring more in the hypothetical grading and 2 scoring less. Of the 28 scoring one or two points for Question 3 (in contrast to zero in the final answer grading), 10 fell short of the full credit of three points due to arithmetic errors and 17 due to not fully understanding what they needed to do. While most arithmetic errors were minor and few, three students made a great number of egregious errors. Two students (Table 5: '2-:understanding') were awarded full credit (three points) in the final answer grading but analysis of the students' working showed that they happened on the correct answer by trial and error rather than using a fool proof strategy (see also Aydin, 2014, for students' 'guess and check' approach to linear dependence). Another noteworthy error was a copy error. In the student's rough work he solved the question correctly but recorded the incorrect information on the final answer sheet. This error resulted in gaining three points in the hypothetical grading.

Question 4 proved a problematic item and contributed more than any of the other items to refining our design principles. Question 4 consists of two parts, Questions 4a and 4b. Each subpart was awarded 3 points in the final answer grading mode. Given a four by three matrix, 4a asked the student to find a basis for the null space of the matrix, while 4b asked for a basis for the column space of the same matrix. The hypothetical grading was structured differently with two points each for correctly reducing the matrix to reduced echelon form (one point if a minor arithmetic error), correctly interpreting the reduced matrix to extract the basis for Null A (one point if Null $A = \text{Span}\{\}$ was given), and for correctly interpreting the reduced matrix to extract the basis for Col A (1 point if Col $A = \text{Span}\{\}$ was given). This grading structure of $2 + 2 + 2$ is structurally different from the final answer structure of $3 + 3$. For this reason, we did not include Question 4 in Table 3 since a breakdown by 4a and 4b would not be possible.

This pair of items was understood to adhere to the design principles of being primarily computational and checkable. In fact, to the trained eye, it could be seen that columns one and three only differ by a change of sign, and columns two and four are the same. Thus, without any calculation it can be seen that columns one and two provide the pivots in reduced echelon form, providing an answer to 4b. The answers are checkable in that if $\mathbf{v}$ is an element of the null space then $A\mathbf{v} = \mathbf{0}$ (so the null space is ostensibly easily checkable) and, it could be argued, since the same process of reduction informs the null space a correct null space implies a correct column space.

The assumptions made in the design of this item proved problematic. Firstly, only if students wrote 'a basis for Null/Col $A = \{(),()\}$' or simply '$\{(),()\}$' would they gain full points, for a or b, respectively. However, since this was a paper proxy of a digital testing, grader discretion played a role here as some graders were more lenient about notational variation. A consensus was reached that if a student erroneously wrote the words 'span' in both answers, and was otherwise completely correct, that students only lost points for one of the sub-questions. A further problem was the non-uniqueness of bases. Three students column reduced instead of row reduced, resulting in a basis that did not look like the grading

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

scheme but was indeed correct. Sangwin & Köcher (2016) refer to the unsuitability of items with non-unique correct answers to a paper test, arguing that sort of question might be more suited to digital testing.

In the hypothetical grading, most students (113 of 124, 91%) were aware that a useful first step was to reduce the matrix. Most of those students (94 of 113, 83%) carried out the reduction process correctly. The greatest source of difficulty was in interpreting the reduced matrix usefully. Difficulties at this point included making no further progress, not interpreting pivot columns correctly, notational errors involving 'span', and including vectors of the wrong dimensions in the bases.

For three points, in Question 5a students were given a three-by-three matrix, informed that it has an eigenvalue of -1 and were asked to determine the corresponding eigenspace. This item fits with the existing design principles as it was quickly checkable. Given the eigenvalue of -1, the corresponding eigenvector v should satisfy $A\mathbf{v} = -\mathbf{v}$. The hypothetical grading allowed one point for A-λI, one point for reduction and one point for interpreting the result to find the eigenspace. If an error is made early on, then the student is unlikely to find anything other than the trivial solution for Null(A-λI) as required. A further half point was possible if there was evidence that the student recognized this situation as incorrect.

The notational confusion with respect to the term 'span' was not as evident in Question 5a as in Question 4 for some reason. As seen in Table 2, in total, 25 students were awarded different grades across the two modes of final answer grading and hypothetical grading. Many of those differences were due to beginning the computation correctly but being unable to interpret the results. Omitting the 'span' from the answer was interpreted by us as an error in understanding not as a careless error, since a finite set is fundamentally different from the span of that finite set. Three students (3+:arithmetic) completed the question perfectly but made a copying error when entering their answer onto the answer sheet for grading. In total, 77 (62.1%) students got full points for this question in both modes of grading and 22 (17.7%) got no points in both modes of grading.

For two points, in Question 5b students were asked to determine other (possibly complex) eigenvalue(s) of the matrix A, which was also asked about in Question 5a. This item fits with our design principles by being checkable as you can substitute back the eigenvalues and find the determinant. The answer to 5b is $\lambda = 1 \pm \mathrm{i}$. The hypothetical grading allowed a half point for subtracting λ down the main diagonal, one point for determining the characteristic polynomial and a half point for determining the roots of the polynomial (other than $\lambda = -1$).

The understanding errors observed were related to not knowing how to find the characteristic equation or in how to interpret its roots. This item was subject to a wide variety of arithmetic errors including not knowing how to factorize a polynomial, losing a solution through incorrect factorizing and not knowing how to use the quadratic root formula properly. As seen in Table 2 one student (2+:arithmetic) completed the item correctly, however, did not copy his answer onto the answer sheet for grading (and hence this was classified in our scheme as a copying error and bundled with arithmetic errors). This student did not complete enough work on the entire test to pass and we assume that he gave up, not bothering to submit his correct result for 5b. In total, 74 (59.7%) students got full grades for both modes of grading for this question and 14 students (11.3%) got zero grades for both modes of grading in this question.

6. Discussion

Running this project and analysing the data at this level of detail has raised four points of interest. First, the practicalities of the 'paper proxy' for final answer digital testing made us aware of grader variation, grader error and the strengths and weaknesses of using pure digital grading. Secondly, Question 4 provided a wealth of information on using null space and column space in final answer testing. Thirdly,

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

the design principle related to answers being checkable came under scrutiny during our analysis, and finally we are forced to confront the practice of allowing students to accrue assessment points through computational manipulation not underpinned by true understanding.

We were surprised at the amount of grader variation in the final answer grading given that the grading scheme was assumed to be very clear: a correct answer is awarded full credit (2 or 3 points) and an incorrect answer zero points. However, some graders were more forgiving of notational carelessness or very minor errors than others. In addition to grader variation, there was also simple grader error, for instance not noticing an incorrect sign. Both grader error and grader variation would disappear were the test to be truly tested digitally rather than by hand as a proxy for digital grading.

When Question 4 was initially designed we did not consider it a particularly problematic item; however, our analysis clearly indicated it to be unsuitable for final answer testing, in particular for our paper proxy for digital testing. First, having two similar items raised the double penalty issue, where students might be penalized twice for a single error—in this case, incorrect use of the term 'span' (see also Ashton *et al.*, 2006). Certain digital testing systems could either penalize twice or the test could have been designed so that students needed only to enter numbers, which would then weaken the items as regards testing whether the students know the difference between a basis and the span of that basis, concepts known to be problematic (Stewart & Thomas, 2010; Griffiths & Shionis, 2021). Secondly, bases are not unique. Graders expecting to see one answer might consider a different correct answer as incorrect. Digital testing would only avoid this issue if the system were able to use a computer algebra system to test for equivalence of bases. Grading work by students that give a variety of correct answers, Sangwin (2019a) agrees that computer algebra systems are ideal, as this saves teachers from making mistakes and a tremendous amount of checking. Thirdly, linear spaces have notation requirements not limited to use of the 'span', such as correct use of set notation. Hand grading allows for flexibility which a digital system might not. A digital system allowing enough flexibility for students to demonstrate their understanding through their choice of syntax runs the risk of the double penalty issue or lack of recognition of unexpected but correct answers. Different digital systems themselves vary in their affordances and constraints. Items with answers that have multiple different notational complications might be ones to avoid in digital grading or a paper-proxy as discussed here.

A key design principle was that either answering an item should be a one step process not susceptible to careless error OR that the answers should be checkable so that any careless errors could be detected and corrected (see also Challis *et al.*, 2003). Question 1 was not checkable by a process other than repeating the calculation; however, the numbers involved in calculating the determinant and the prevalence of zeroes means that we decided that this was an item testing one single step process not susceptible to arithmetic error. Question 3 was checkable by considering the coordinates with respect to the basis; however, that process involves several steps. Furthermore, this particular case involved fractional coordinates which increases potential for error and time taken for the checking process. We do consider this item satisfactorily designed but it could have been better with simpler, non-fraction, coordinates.

Question 4 had some parts that were more easily checked than others. A vector $\mathbf{v}$ in the basis of the Null space of A satisfies $A\mathbf{v} = \mathbf{0}$. However, if Null A has dimension two (as this one does) and a student only found one vector then this check does not alert them to the missing vector. The only fast way to check Col A is by checking Null A and concluding Col A must be correct because the same computational process gets you these two results. This approach is not satisfactory. The weak adherence to the checkability design principle underscores the unsuitability of this item, in final answer grading. Question 5a was quickly checkable. Question 5b is slightly more laborious to check, requiring computing a determinant. For ease of checking, which we consider a key design principle, perhaps final answer grading for eigenvectors should only include two by two matrices not three by three.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

Many of the instances of students achieving more points on the hypothetical grading than in the final answer grading were from being given points for setting up a problem, perhaps doing some calculation, but then not following through with correct interpretation of results. For example, two students gained 7.5 points each in the hypothetical partial credit grading and yet still did not achieve a passing grade. These students did not understand many fundamental concepts. They gained points through minor actions in several questions but repeatedly fell short of displaying true understanding. If what we truly value in linear algebra is conceptual understanding and interpretation of information, then by awarding points for setting up matrices without follow through we are sending entirely the wrong message. A common complaint from the students is that they would have been awarded one or two points in partial credit grading and thus final answer grading is unfair, a point also raised by Challis *et al.* (2003). For instance, in Question 3 in several cases, a student would set up and reduce the augmented matrix and then stop, not knowing how to continue. It could be argued that without that final step of interpreting the result the preceding steps are not worthy of credit. The inflexibility of final answer grading, giving no credit for half-answers lacking conceptual follow through, can be considered an advantage by teachers but a disadvantage by students.

7. Conclusion and revised design principles

Grader error and variation would disappear if the test were truly tested digitally rather than using our paper-proxy system. For reliable grading consistency is important when multiple people are grading the same test; however, one category of grader variation occurred due to the perceived unfairness of penalizing the same error twice as we had in the two basis items (Question 4) contributing to refining the design principles. We conclude that true digital grading is more reliable than hand grading as a proxy for digital grading but that goes hand in hand with a well-designed test that does not feature the double penalty issue.

Our analysis of Question 4 provided us with two important new design principles for digital final answer testing. First, do not include two items that might result in penalizing the same error twice. Secondly, when an item has (infinitely) many correct answers only include it if the digital system has a means of testing for equivalence.

We have added the word ‘quickly’ to our design principle related to checking. If it takes as long to check the answer as it did to answer the question, it renders the item unsuitable unless unlimited time has been made available for the writing of the test. Similarly, the checking process itself should be relatively immune to careless error, a characteristic not shown by Question 3, for instance.

Refined set of design principles:

- Limit the number of points assigned to an item to a maximum of three.
- An item should be a one step process not susceptible to careless error OR the item should be quickly checkable and students should have had the opportunity to learn how to do this.
- The algebra involved should be minor, testing the core concept of the question, rather than algebraic manipulation. Consider avoiding fractions.
- Avoid items where the same error could be penalized twice.
- Avoid items with notational complexities where (were the item to be hand graded) grader opinion might differ
- If infinitely many answers are correct (for example, a basis of a subspace) the testing procedure must allow for recognition of any correct answer.
- As the learning goals that can be tested in this way are limited, a maximum of 50% of the grade should be based on final answer questions.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

In a recent survey of educational literature on teaching and learning linear algebra (Stewart *et al.*, 2019) very little pertained to assessment and none to anything akin to 'final answer' testing—a concerning omission. Sangwin (2019a) reports 'modest success' (p. 7) in the implementation of automated assessment of a mock exam for linear algebra. He raises the concern of confidence in examination validity, certainly a factor in final answer testing. We have put the design principles discussed in this article effectively and satisfactorily to the test in more recent linear algebra assessment. Our article contributes to the conversation on use-inspired research into computer-aided assessment of university mathematics (Kinnear *et al.*, 2020) and of linear algebra specifically. We hope that this work can be of help to our colleagues in higher education engaged in exploring the potential of (digital) final answer testing.

Acknowledgments

We are grateful to Steven Horstink for carrying out the grading and to Gaby Ong for the inter-rater reliability checking.

References

Aarts, H. (2018) A hybrid test for mathematics. Unpublished report for Senior University Teaching Qualification, University of Twente. Available from the authors upon request. Poster presentation accessed via https://bit.ly/3wlKfqI.

Ashton, H. S., Beevers, C. E., Korabinski, A. A. & Youngson, M. A. (2006) Incorporating partial credit in computer-aided assessment of mathematics in secondary education. *Br. J. Educ. Technol.*, 37, 93–119.

Aydin, S. (2014) Using example generation to explore students' understanding of the concepts of linear dependence/independence in linear algebra. *Int. J. Math. Educ. Sci. Technol.*, 45, 813–826.

Broughton, S. J., Robinson, C. L. & Hernandez-Martinez, P. (2013) Lecturers' perspectives on the use of a mathematics-based computer-aided assessment system. *Teach. Math. Its Appl.*, 32, 88–94.

Challis, N., Houston, K. & Stirling, D. (2003) Supporting good practice in assessment in mathematics, statistics and operational research.

Griffiths, B. J. & Shionis, S. (2021) Student attitudes towards linear algebra: an attempt to roll back the fog. *Teach. Math. its Appl.*, 40, 182–189.

Kinnear, G., Jones, I. & Sangwin, C. (2020) Towards a shared research agenda for computer-aided assessment of university mathematics. *Math. Educ. Digital Age*, 377–384.

Koomen, M. & Zoanetti, N. (2018) Strategic planning tools for large-scale technology-based assessments. *Assess. Educ. Princip. Policy Pract.*, 25, 200–223.

Lawson, D. (2002) Computer-aided assessment in mathematics: panacea or propaganda? *Int. J. Innov. Sci. Math. Educ.*, 9, 1–12.

Lochner, A. J. (2019) Summative digital testing in undergraduate mathematics: to what extent can digital testing be included in first year calculus summative exams, for engineering students? Accessed via https://essay.utwente.nl/77167/1/Lochner_MA_EST.pdf (21 May 2021).

Martins, S. G. (2018) A study of the application of weekly online quizzes in two courses of mathematics for engineering students—is it a fair and effective strategy to increase students' learning? *Int. J. Innov. Sci. Math. Educ.*, 26.

Mcguire, G. R., Youngson, M. A., Korabinski, A. A. & Mcmillan, D. (2002) Partial credit in mathematics exams-a comparison of traditional and CAA exams. *Proceedings of the 6th CAA Conference*. Loughborough: Loughborough University.

McKenney, S. & Reeves, T. C. (2018) *Conducting Educational Design Research*. London: Routledge.

Moccozet, L., Benkacem, O., Tardy, C., Berisha, E., Trindade, R. T. & Bïrgi, P. Y. (2018) A versatile and flexible framework for e-assessment in higher-education. *2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET)*. Olhao, Portugal: IEEE, pp. 1–6.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

Rane, V. & Mackenzie, C. A. (2020) Evaluating students with online testing modules in engineering economics: a comparision of student performance with online testing and with traditional assessments. *Eng. Econ.*, 65, 213–235.

Robinson, C. L., Hernandez-Martinez, P. & Broughton, S. (2012) Mathematics lecturers' practice and perception of computer-aided assessment. *Mapping University Mathematics Assessment Practices* (P. Iannone & A. Simpson eds), Norwich: University of East Anglia, pp. 105–117.

Sangwin, C. (2019a) Developing and evaluating an online linear algebra examination for university mathematics. *Eleventh Congress of the European Society for Research in Mathematics Education (No. 15)*. Utrecht, Netherlands: Freudenthal Group; Freudenthal Institute; ERME.

Sangwin, C. (2019b) Reasoning by equivalence: the potential contribution of an automatic proof checker. *Proof Technology in Mathematics Research and Teaching. Mathematics Education in the Digital Era*, vol. 14 (G. Hanna, D. A. Reid & M. de Villiers eds). Cham, Switzerland: SpringerLink, pp. 313–330 https://doi.org/10.1007/978-3-030-28483-1_15.

Sangwin, C. J. & Köcher, N. (2016) Automation of mathematics examinations. *Comput. Educ.*, 94, 215–227.

Stewart, S. & Thomas, M. O. (2010) Student learning of basis, span and linear independence in linear algebra. *Int. J. Math. Educ. Sci. Technol.*, 41, 173–188.

Stewart, S., Andrews-Larson, C. & Zandieh, M. (2019) Linear algebra teaching and learning: themes from recent research and evolving research priorities. *ZDM*, 51, 1017–1030.

Yerushalmy, M., Nagari-Haddif, G. & Olsher, S. (2017) Design of tasks for online assessment that supports understanding of students' conceptions. *ZDM*, 49, 701–716.

Appendix

Items 1, 3, 4 and 5 were 'final answer' items and Items 2, 6 and 7 were graded on students' fully worked solutions.

Item 1.
Given is the matrix

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}.$$

a) Determine $\det(A)$.
b) Determine $\det(A^{-1})$.

Item 3.
Consider a vector space and four vectors in $\mathbb{R}^3$:

$$\mathcal{T} = \text{Span}\left\{\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}\right\} \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -1 \\ 2 \\ 4 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 1 \\ 6 \\ 3 \end{pmatrix}, \mathbf{v}_4 = \begin{pmatrix} 1 \\ 5 \\ 4 \end{pmatrix}$$

Indicate for each of these four vectors whether they are an element of $\mathcal{T}$ or not.

Item 4.
Given is the matrix

$$A = \begin{pmatrix} 1 & 2 & -1 & 2 \\ 1 & 1 & -1 & 1 \\ -1 & 0 & 1 & 0 \end{pmatrix}.$$

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022

a) Determine a basis for Null A.
b) Determine a basis for Col A.

Item 5.
The matrix A is given by

$$A = \begin{pmatrix} 1 & -3 & -1 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

a) A has eigenvalue -1 (you do not have to prove this)
Determine the corresponding eigenspace.
b) Determine the other (possibly complex) eigenvalue(s) of A.

Alisa Veale (nee Lochner) (alisa.veale@utwente.nl): Alisa has a master's in educational science and technology at the University of Twente and a bachelor's in mathematics and computer science from the University of Rhodes. She also has experience in teaching High School Mathematics in South Africa for a few years. From 2019, she continued to work at the University of Twente for the faculty of EEMCS regarding educational support, e-learning and testing.

Tracy Craig (t.s.craig@utwente.nl) is a lecturer of mathematics at the University of Twente, teaching predominantly calculus and linear algebra to students in technical programmes. Before 2018 she was situated at the University of Cape Town, South Africa. Her research is practice-led in the areas of mathematics education and engineering education, with particular interest in the teaching and learning of linear algebra, vectors and vector calculus, as well as in the Twente Educational Model.

Downloaded from https://academic.oup.com/teamat/advance-article/doi/10.1093/teamat/hrac002/6545982 by Universiteit Twente user on 14 March 2022