CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

The feasibility of on-screen mocks in maths and science

Research Report

Joanna Williamson for the Digital High Stakes Assessment Programme

25 January 2023

Author contact details:

Joanna Williamson
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

joanna.williamson@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division
If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Williamson, J. (2023). *The feasibility of on-screen mocks in maths and science*. Cambridge University Press & Assessment.

Acknowledgments
Sincere thanks go to all the Cambridge colleagues who shared their experiences and insights as part of this project, particularly Dan Bray, Stuart Briner, Phil Cowen, Markus Hextall, Sarah Hughes, Ami Jones, Heather Mahy, Sanjay Mistry, Ross Roberton, Ed Sutton, and Brooke Wyatt. I would also like to thank Natalia Harvey for administrative help preparing the report.

Summary

The Cambridge Digital Mocks Service offers on-screen assessments created by transferring paper-based (PB) exams into a digital test platform. This feasibility project considered the validity of extending the Digital Mocks Service to maths and science qualifications, thinking about (i) the outcomes that are likely when maths and science items are 'lifted and shifted' into digital formats, (ii) the intended purposes of the Digital Mocks Service assessments, and (iii) how far maths and science assessments in the Digital Mocks Service could meet these intended purposes. The project drew on published research, the knowledge and experience of colleagues, and findings from in-house trials.

The key findings were:

- Shifting a PB item to a digital format can alter both the mathematical and scientific response activity that items elicit, and the expression of candidates' ideas. These effects can compound one another, since where is a lack of a sufficiently seamless way to express mathematical and scientific ideas within the digital environment, this can be one of the ways in which the digital test platform alters response activity.
- Although familiarity is a partial explanation, there are inherent challenges in inputting mathematical and scientific notation in digital environments. It is important to recognise that some digital assessment solutions can be both unfamiliar and objectively cumbersome.
- Items can be adapted to minimise 'difficult' inputs in their on-screen format, but this is likely to alter the item demand or constructs assessed, or both.
- The use of tablets instead of keyboard and mouse input reduces, but does not remove, the difficulties of expressing mathematical and scientific ideas. It also introduces new challenges in terms of screen space, device comparability, and device availability.
- There can be significant differences between candidate performance on PB and digital items, even for items that were apparently straightforward to shift to an on-screen format.
- Unless on-screen assessment of maths and science takes place using the same tools and methods used in teaching and learning, there is a high risk of assessing ICT literacy rather than mathematical and scientific skills and knowledge.
- Options for maths and science assessments in the Digital Mocks Service include omitting difficult-to-shift items, retaining these items in altered forms, replacing them with alternative items that assess similar content, or restricting the assessments to classrooms where teaching and learning takes place using the same digital platform as the Digital Mocks Service (or a very similar one). None of these options would offer students an authentic 'practice run' for their final PB exams or produce mock grades comparable to those from PB exams. However, other purposes of a Digital Mocks Service assessment could still be met, namely providing an insight into student strengths and weaknesses, providing an additional assessment opportunity, and providing an external assessment opportunity.

Introduction

High-stakes assessments for school-based qualifications such as GCSEs and A levels remain for the moment largely paper-based. There is, however, increasing interest in the assessment innovations made possible by digital technology, and increasing demand for digital learning and assessment solutions even while end-of-course assessments remain paper-based. The Cambridge Digital Mocks Service was set up to meet customer needs by offering on-screen assessments as practice for current (paper-based) exams. So far, the Digital Mocks Service has developed and trialled assessments in GCSE Computer Science, International AS level History and IGCSE English as a Second Language. In each case, the assessment consisted of an entire PB exam that had been migrated to the digital test platform. The assessments were marked by Cambridge examiners, and marks (but not grades) reported back to schools. Exploring how teachers and schools would want and expect to use a product such as the Digital Mocks Service was one of the aims of the trials.

The question investigated in this feasibility project was about the validity of extending the Digital Mocks Service to maths and science qualifications. Could the Digital Mocks Service meet its intended purposes for maths and science exams? Specifically, how feasible is it to include maths and science qualifications in the Digital Mocks Service, given what we know about:

- the impact of migrating maths and science paper tests to screen on validity
- teachers' preferences around having the papers marked for them by Cambridge or marking the papers themselves
- the purpose of the Digital Mocks Service?

In this short report, the first and largest section lays out the outcomes that are likely when maths and science items are 'lifted and shifted' into on-screen formats, based on the existing academic literature, and the knowledge and experience of colleagues. The second section considers the intended purposes of the Digital Mocks Service assessments, what maths and science assessments in the Digital Mocks Service might look like, and how far these assessments would meet the intended purposes of the Digital Mocks Service.

The specific focus of this short report on the Digital Mocks Service is important. There are of course related questions about the digital assessment of maths and science more generally (Appendix 1), and answering these involves thinking about fundamental questions such as the purposes of maths and science education at school level. These questions were out of scope for the current project, although the evidence assembled in this report has relevance.

1. What outcomes are likely when existing PB maths and science items are shifted into digital tests?

From a practical perspective, it is helpful to consider the outcomes of 'lifting and shifting' maths and science items in two stages:

1. The extent of alteration required to 'lift and shift' an item into a functioning digital version in the chosen digital platform, highlighting ways in which we might (theoretically) expect validity and comparability with the PB original to be affected.
2. Assessment data and/or experimental evidence on the validity of the resulting digital item, and its comparability with the PB item, in terms of the response strategy elicited, the construct assessed, facility level, accessibility to candidates, discrimination properties and lack of construct-irrelevant variance.

The extent of comparability between PB and digital versions matters because the items being transferred from PB to digital formats were developed for PB assessment, and because the high-stakes end-of-course assessments of the maths and science qualifications that could be added to the Digital Mocks Service are currently still PB exams. The point here is not to argue the need for comparability between PB and digital maths and science assessment items in general.

Our evidence on the likely outcomes of lifting and shifting maths and science items comes from published research, trials and pilot studies, and operational experience. In summary, the evidence shows that lifting and shifting maths and science items produces variable results: it is relatively straightforward for some items and not at all straightforward for others, and the digital versions of items are sometimes but not always comparable to their PB counterparts. The relevant factors include the characteristics of the original PB item, the design and capabilities of the digital testing platform, the capabilities of the intended device(s), and the population of intended test-takers.

1.1 Lifting and shifting maths and science items

The minimum level of alteration we can probably imagine in the shift from PB to digital would be for a short multiple-choice item (MCQ) that fits easily onto a single page and a single screen, with a response format of an arbitrary mark (e.g., a tick in a box) – in the PB version made by hand, and in the digital version with a click or tap. For items such as these, confidence in the comparability of PB and digital item versions might be quite high – and numerous studies have indeed demonstrated comparability in particular testing contexts. Two things complicate the picture: firstly, there are also studies showing significant differences according to testing mode, even for items that were apparently straightforward to shift (e.g., Fishbein et al., 2018). Secondly, although the type of multiple-choice item described above is common in some tests, it accounts for only a small proportion of marks in maths and science GCSEs, IGCSEs and A levels. For longer constructed response items, we have less published experimental evidence, and the evidence we do have suggests that mode effects may increase for such items.

Many aspects of the lifting and shifting of maths and science items are shared with other subjects: the need to manage screen 'real estate' carefully, for instance, to ensure that candidates see what is necessary. We focus first, however, on two characteristics that affect maths and science particularly strongly, and set them apart from other subjects:

1. The expression of mathematical and scientific ideas is often not in words
2. Many items require candidates to **do** things, and for aspects of the 'doing' or 'working' to be captured

These characteristics mean that shifting a PB item to a digital format poses a potential double threat to validity: it has the potential to both alter the mathematical and scientific response activity that items elicit (e.g., the use of mathematical modelling, trial and error, particular calculations or algorithms, sketching, annotation of diagrams, mental calculation and reasoning, written calculation and reasoning, guesswork, and estimation), and alter the expression of candidates' ideas. These effects may also compound one another, since the lack of a sufficiently seamless way to express mathematical and scientific ideas within the digital environment can be one of the ways in which the digital test platform alters response activity. For instance, if inputting steps of algebraic manipulation is cumbersome, a candidate may decide to reduce or omit the writing down of intermediate steps, and rely more heavily on mental operations.

1.1.1 Expressing ideas in maths and science

The expression of mathematical and scientific ideas requires frequent use of non-standard characters (e.g., Greek letters, a degree symbol), special formatting (at a minimum, subscripts, superscripts, and fraction notation) and drawing. The difficulty with all of these is that they are not easy or seamless to input using a keyboard and mouse. Computers equipped with a keyboard and mouse are the digital devices that are most readily available in school settings, however, and are considered the most appropriate for many assessments – where longer responses are likely to require extensive typing (Ofqual, 2020). While some extended responses in maths and science assessments do require writing (of words) (e.g., a detailed description of a biological process), extended responses in maths and physics particularly are more likely to require demonstration of calculations, and the expression of ideas in mathematical and scientific notation.

Many studies emphasise lack of student familiarity with the digital test platform as an explanatory factor for lower candidate performance on digital versions of maths and science items (e.g., Fishbein et al., 2018), but there are also inherent challenges in inputting mathematical and scientific notation in digital environments. That is, we must acknowledge that some digital assessment solutions can be both unfamiliar *and* objectively cumbersome.

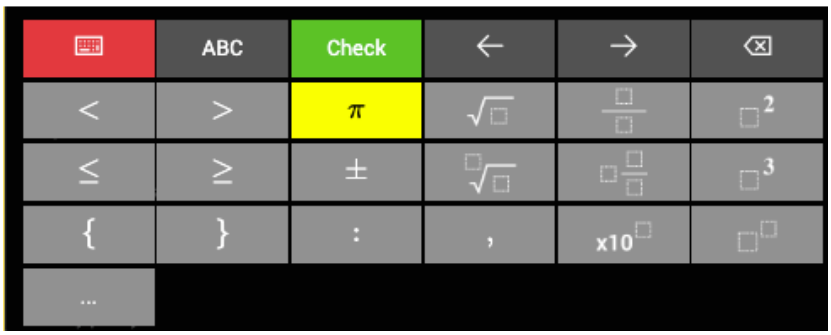
Unlike typing words, typing mathematical and scientific notation doesn't get extensively practised outside of maths and science contexts, and at school level is usually not extensively practised even within maths and science contexts. To assign all the difficulty to practice, however, would be to miss the point – and the history of mathematical writing and typesetting can highlight more precisely what is going on. Mathematics is a language, and "its particular need to express complex ideas in concise ways, has resulted in an especially productive writing system" (Mills & Hudson, 2007, p. 6). At the same time, "Reducing the ideas of mathematicians and scientists to the very solid form of typeset text has been, and

remains, a challenge for authors, editors and typographers" (p. 6) – or, in the blunter terms of Backhouse et al. (1997), "fraught with difficulty". Mathematics was historically and remains "the material most feared" (p. 10) in typesetting because of the combination of the "extraordinary diversity of individual characters" and the fact that "unlike conventional text, mathematics is not linear in its construction: an equation depends on a two-dimensional arrangement" (Mills & Hudson, 2007, p. 12). The struggles of professional mathematicians and scientists have led to extraordinary innovation, such as the TeX language (later LaTeX) invented by Donald Knuth, which can produce excellent outcomes, but with a still-high cost (LaTeX is considered difficult to use even by many professional users).

A common solution to the difficulties of mathematical and scientific input in digital assessment is to provide on-screen equation editors or menus that allow students to insert mathematical symbols and layouts such as fraction notation and exponentials (e.g., Figure 1) using point and click. The drawback of these systems is that they are non-standardised, slower than typing, and cumbersome in comparison with handwriting (Ofqual, 2020, pp. 14-15). The prospect of requiring students to input multiple lines of mathematical reasoning (e.g., Figure 2, Figure 3) in this way within an assessment scenario is somewhat concerning: the user must mentally plan each line of working, work out the symbols required and the order in which to click to achieve the correct arrangement (which is non-trivial, since a single line of working is itself not linear in construction), then repeat the planned clicking and typing of digits many times. Colleagues who have experience of typesetting mathematical working for items such as those in Figure 2 and Figure 3 (e.g., for use in teaching resources) will be aware of the frustration that can be induced, even when the content is mastered, there is no time pressure, and the stakes are low. For students working at the boundaries of their knowledge and understanding, there is a clear threat to assessment validity.

a)

b)



c)

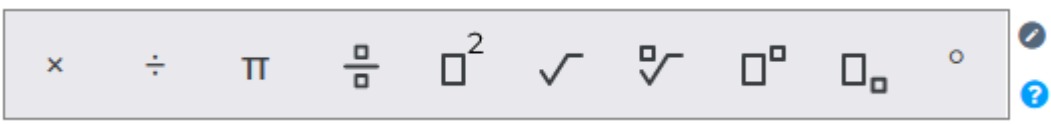


Figure 1: Three examples of on-screen equation editors from commercially-available learning and assessment platforms.

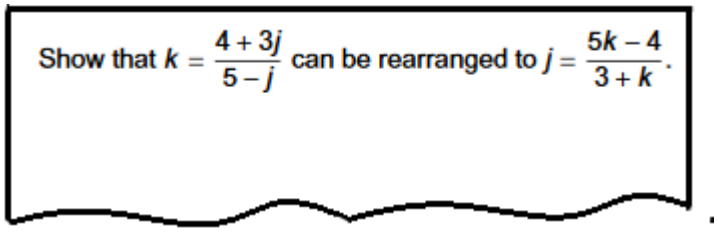


Figure 2: Example of item requiring multiple lines of algebraic manipulation.

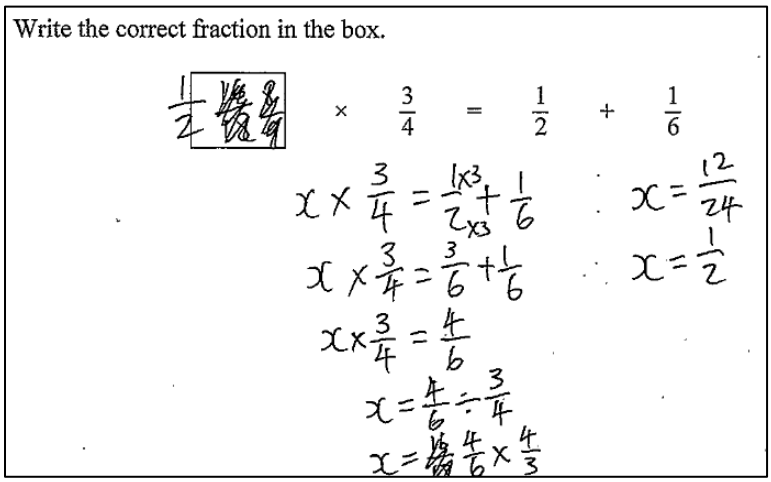


Figure 3: Student example showing multiple lines of working.

Cambridge colleagues have observed the difficulties faced by students in correctly inputting scientific responses in digital assessments. In trials of science tests, for example, some teachers observed that most of their students were unable to correctly input several answers, resulting in the need for the teacher to override the auto-marking facility and mark responses based on what the students had 'meant' rather than what they had actually input. The difficulty is not restricted to complex or advanced items; in secondary science, for instance, students are learning the meaning of the notation H_2O for water, and it is important that this is different from H2O, H_20, H20, H^2O, ho2, and any other similar arrangements.

The use of a tablet instead of keyboard and mouse presents many advantages, as students can handwrite responses, and thus sidestep the limitations of keyboard and mouse input. A recent study by Aspiranti et al. (2020), however, demonstrates some useful insights into the nuances of working on tablets. In particular, "All six students indicated that they did not like the stylus because they could not rest their hand or arm on the tablet the way they typically would when writing on paper. When this occurred, the tablet would jump from the stylus to where the student's arm lay, making marks across the paper." (Aspiranti et al., 2020, p. 461). In addition, half of the students reported that "they did not like the squishy ball at the end of the stylus because it was not the same as actually writing" (p. 461). All students performed better on the PB items than on the tablet-input and keyboard-input digital versions of items. This study was very small in scale, and it would not be appropriate to generalise from the elementary school sample to older age-groups and different countries without further research. It offers, however, a powerful reminder of how much the details of tools and user experience matter. Despite offering a digital solution in which students were theoretically able to handwrite their answers, the need to 'hover' their hand and arm in the

air whilst writing was perceived as a noteworthy obstacle, and attempts to write naturally induced frustration at the unwanted marks that appeared. The use of a tablet instead of a laptop or desktop computer also introduces other considerations: variation in device quality and size, pop-up keyboards, reduced screen space for assessment elements, and the fact that schools may not have sufficient numbers of devices of a size and quality sufficient for a Digital Mocks Service assessment. A teacher in the Digital Mocks Service Computer Science trial, for instance, remarked that it would be interesting to trial the assessments on tablet, but that "we haven't got anywhere near enough of those".

There is general agreement that valid high-stakes digital assessment of mathematics can only take place when students are accustomed to using the same tools in their mathematics learning. Drijvers (2019), for instance, is emphatic:

> *"To foster test validity, it is crucial that these mathematical tools in the assessment player are similar to the tools that students use in the preceding teaching and learning. This will avoid test artifacts that relate to the user interface of these tools, and to their limitations and constraints. Once students are familiar with these tools, we can avoid assessing students' ICT literacies rather than their mathematical knowledge. If students are not prepared for the use of the digital tools during the test, test validity is already threatened …" (Drijvers, 2019, pp. 61-62).*

This point has been noted by Ofqual as one of the key barriers to high-stakes digital assessment. Specifically, while "familiarity" is frequently cited as a means to overcome the difficulties of input in digital maths and science assessment, this familiarity "depends on full integration of teaching and learning methods (which may vary depending on device type, availability and other factors) into an assessment system, or replicating the assessment system into a variety of operating systems" (Ofqual, 2020, p. 15). Furthermore, there would need to be cross-subject consistency, with "mathematics and physics using different means to capture mathematical notation" noted as a particular problem to avoid (p. 15).

In England, coordination between different awarding organisations (AOs) would also seem necessary, since schools may use different AOs for their maths and science qualifications. In an international context, similarly, it would be necessary to consider consistency between providers (e.g., national and international assessment providers) for schools using more than one. A further point that Drijvers notes is that creating a sufficiently capable environment for this mathematics assessment requires perhaps surprisingly sophisticated tools. Specifically, "the experiences in the Dutch diagnostic test development have shown that, even for relatively "easy" mathematics, we need "hard" tools such as computer algebra systems and dynamic geometry systems" (Drijvers, 2019, p. 61).

The Digital Mocks Service will help Cambridge to learn and prepare for high stakes onscreen assessments in future. A question for the immediate term, though, is how far the validity requirements set out above could acceptably be relaxed, for the purposes of low-stakes assessment. Despite lower stakes, it is still fundamentally important for an assessment offered as a maths assessment to "avoid assessing students' ICT literacies rather than their mathematical knowledge" (Drijvers, 2019, p. 43).

1.1.2 Minimising difficult inputs

An adaptation sometimes made when lifting and shifting PB items into digital formats is to introduce scaffolding for responses, for instance by providing the necessary measurement units, equation form (fill-in-the-blanks), or answer spaces in correct locations on a diagram. The major advantage of such scaffolding is that responses are simplified to typing in numbers, removing the challenges of mathematical and scientific notation mentioned previously. Quite clearly, however, the scaffolding changes the assessment item. An example of such scaffolding is shown in Figure 4. This item might still be considered a worthwhile item, but it does not assess the same constructs as the original item: most notably, the problem-solving and planning aspects of the original item have been lost. More generally, Green and Hughes (2022) applied the CRAS framework of demand (Complexity, Resources, Abstraction, Strategy – Task and Strategy – Response) to demonstrate how item demand is likely to reduce in terms of strategy (both task and response) when items are re-written into more objective item types.

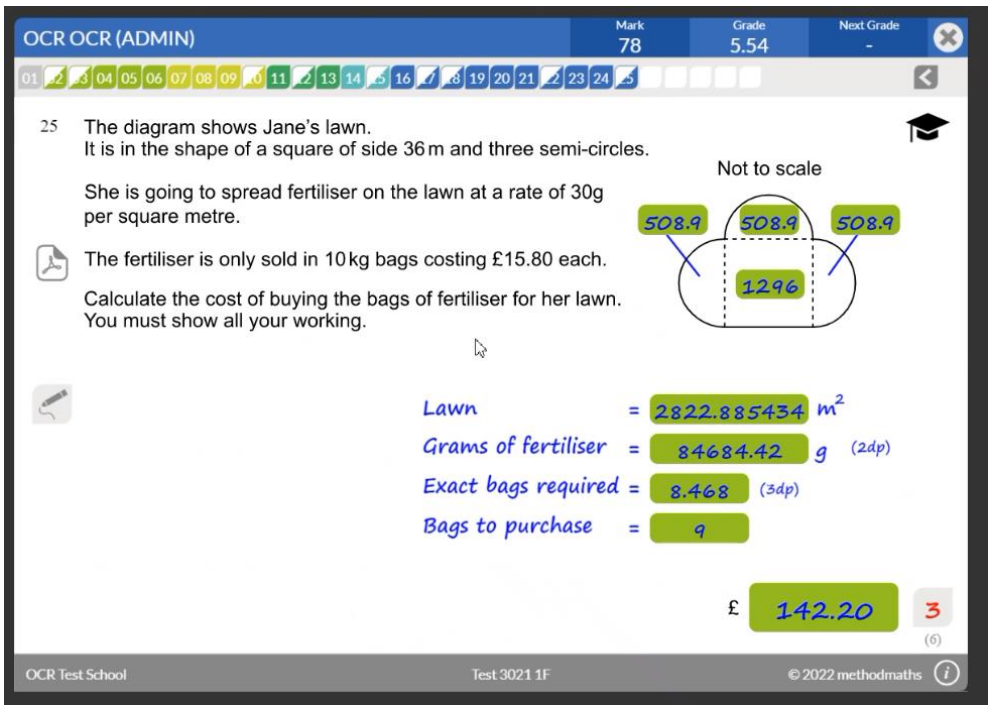


Figure 4: Example of scaffolded response space.

Another common way to avoid extensive constructed inputs is through using multiple-choice items. Multiple-choice questions (MCQs) can be used successfully in maths assessment, but the research highlights the need for very careful item design to achieve valid assessment of the mathematics constructs targeted, and it certainly cannot be assumed that parallel constructed response and MCQ 'versions' of an item measure the same mathematical knowledge and skills (Shepard, 2008). If the intention is to elicit a particular calculation to solve a problem for instance, the item and response options need to be designed in such a way that candidates cannot use back-substitution in a trial-and-error strategy to find the correct solution – asking candidates to select the answer option closest to the true result is one possible method. An alternative is to accept the use of back substitution, and factor in this strategy in designing the complexity and demand of the item. A trial of digital assessment for A level maths by OCR confirmed that it was necessary to write MCQs from

scratch rather than convert constructed response PB items in order to achieve valid MCQ items at this level. Under the current approach of the Digital Mocks Service, however, the goal is explicitly to re-use PB items.

1.1.3 Recording of 'working'

Many mark schemes in maths and science award marks for demonstration of 'working', and/or give partial credit for answers that carry forward an initial error in subsequent correct steps. To add maths and science assessments to the Digital Mocks Service using the current 'lifting and shifting' approach, capturing student working would be essential.

In PB formats, a challenge of assessing working is encouraging candidates to write down their working, rather than just a final answer, and this challenge is amplified in a digital test format. Johnson and Green (2006, p. 25), for instance, observed differences in response behaviour that seemed to be associated with the ease of writing down working:

> *"There were three questions where students performed significantly better on paper than on the computer and here, performances appeared to be influenced by scratch paper. For these questions on the computer students were less likely to show their work. … Restating the point, if the student thinks the calculation is easy enough he/she will do it mentally from the screen. If the question is already on paper it is more natural, due to familiarity, and takes less effort for the student to use written methods to support his/her thinking. It might be speculated that this is where mode may most clearly influence a student's strategy choice."*

Table 1 lists four possible solutions to capturing working in on-screen tests, and the associated advantages and disadvantages.

Table 1: Ways to capture working in a digital test platform.

Potential solution	Pros	Cons
Working remains on paper or in candidate's head; candidate is asked to transfer the working to a response box (e.g., typing the steps of a calculation)	Candidate is not required to change working-out method Candidate can practice same methods applicable in a PB exam	Time and effort cost of transfer may be large (e.g., for steps of algebraic manipulation) Risk of self-censorship in the transfer Risk of slips in the transfer Unknown impact on candidate cognitive strategies, and on teaching and learning
Working takes place in the digital environment via use of a tablet; candidates handwrite into the response space using stylus/finger	Candidate has same options as on paper: freestyle writing and drawing, and option to cross/rub out Can practice same methods applicable in a PB exam No need to transfer working, saving time, effort and risk	Tablet required Variation in device quality and device familiarity could each create between-candidate variation (e.g., via frustration for those with poor stylus) Candidates might reject digital environment -> working then to be transferred (or not), with attendant risks Writing with a stylus differs from writing with a pen/pencil on paper Unknown impact on candidate cognitive strategies, and on teaching and learning

Working takes place in the digital environment; candidate is asked to type working	Possible on wide variety of devices – anything with a keyboard No need to transfer working, saving time, effort and risk	Difficult or impossible to 'force' candidate to work this way – candidate may switch method (e.g., increase reliance on mental maths) Arguably only efficient when candidate knows model answer – liable to be cumbersome and frustrating for others Overall impact on candidate cognitive strategies, and on teaching and learning, remains unknown
Working takes place in the digital environment and working is captured by recording keystrokes	An appropriate tool may be easily available (e.g., on-screen calculator) No need to transfer working, saving time, effort and risk	Current platforms don't appear to offer this - would keystroke data for method marks require auto-marking to be viable? Invalid if candidate does not carry out working in the digital environment (e.g., can't, won't, decides too cumbersome) Invasive? Can the candidate decide what is submitted? No obvious way to 'cross out' an abandoned attempt. Unknown impact on candidate cognitive strategies, and on teaching and learning

1.1.4 Drawing, annotating and constructing

In maths and science items that are shifted on-screen, the identification of points can be replicated straightforwardly, using a click instead of manual X or drawing of an arrow.

The use of physical measurement tools needs to be either omitted, or replicated in on-screen versions of physical tools. This can appear inauthentic, and confusing if (for example) screen size and resolution means that an on-screen 'ruler' appears at odds with its own measurements. Other potential difficulties include accurately manipulating the on-screen tool, and ease of reading off measurements (Johnson & Green, 2006, p. 22). Where these problems are avoided, however, items that require physical tools in their PB form can transfer well, as in the construction example shown in Figure 5. Provided that students had seen this environment and learned how to manipulate the points and lines, valid assessment of the principles of construction would be possible. Some students in mode effects research have also expressed preferences for on-screen tools, for instance, that an on-screen protractor "wobbled less" than a manual protractor on paper, and that it was impossible to use upside-down by accident, as the digital platform presented the tool correctly oriented by default (Johnson & Green, 2006, p. 22).

Digital test platforms can alter student response behaviour by preventing the physical interactions that are possible with PB items. Many items requiring measurement tools or construction techniques have a strong spatial element, and Johnson and Green (2006) observed mode-related effects such as students preferring to rotate the paper in a PB angle measurement question, and craning their necks while attempting the digital version. Overall, the students in Johnson and Green's sample expressed a preference for the PB versions of

shape, space and measurement items, over the digital versions (2006, p. 22), in agreement with earlier work by Greenwood et al. (2000).

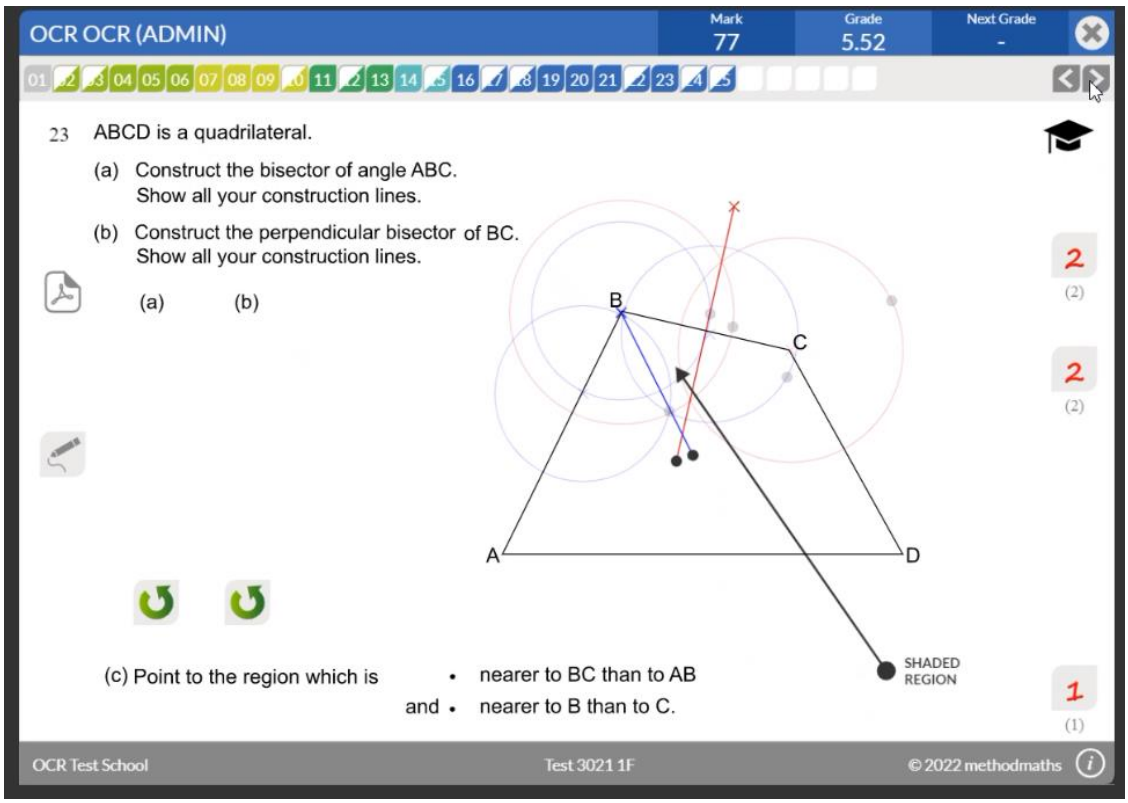


Figure 5: Example of geometry construction item.

1.2 The validity and comparability of lifted and shifted items

Many studies on the validity of digital maths and science assessment evaluate innovative items, that have been designed to capitalise on the affordances of a digital assessment environment. Whilst these studies are often inspiring, they are of limited use for the question of the feasibility of the Digital Mocks Service, where the intention is to re-use items written for the PB format.

A key source of evidence on the performance of 'lifted and shifted' items is mode effects research. In some cases, this research was carried out in order to support the 'lifting and shifting' of existing assessments, exactly as intended in the Digital Mocks Service; in other cases, the lifting and shifting of items was carried out in order to experimentally isolate the effects of assessment mode as far as possible. Either way, such mode effects research offers the most detailed and relevant evidence on how lifted and shifted items in maths and science perform.

1.2.1 International research evidence

A comprehensive literature review of recent research on mode effects was carried out by psychometricians at ACT (Arthur et al., 2020). In maths, the review identified five studies showing that candidates scored more highly on PB tests than their digital equivalents, and five studies showing comparability between PB and digital tests. A study of mode effects in

PISA 2012, meanwhile, showed different results by country: mode effects in favour of PB tests in some countries, in favour of digital tests in other countries, and comparability between PB and digital tests in other countries (Jerrim, 2016). In Science, the ACT review identified three studies showing the candidates scored more highly on PB tests than their digital equivalents, and four studies showing comparability between PB and digital tests. The high-level view confirms that we can't assume comparability, and that mode effects are sensitive to context (items, candidates, devices, platforms), but does not offer much insight into this report's central feasibility question.

Two strands of research particularly worth highlighting are those relating to shifting PISA and TIMSS items from PB to digital formats. These research projects involved very large samples of test takers; the samples of test takers were diverse (from different education contexts and with different classroom norms); there was a focus on maths and science items; and the research sought highly detailed information about the equivalence of PB and digital items (especially, for which items and for which students) rather than considering only the average effects.

TIMSS

In preparation for launching eTIMSS, researchers for TIMSS carried out extensive experimental comparisons of PB and digital versions of maths items (Fishbein et al., 2018)[1]. TIMSS experts classified each item according to its hypothesized likelihood of being "strongly equivalent" or "invariant" between PB and digital test formats. This pre-test classification was based on item features identified in prior research and pilot studies as possible sources of mode effects:

- Differences in presentation between the PB and digital item format, such as essential formatting changes
- Complex graphs or diagrams, or heavy reading loads, which may require greater cognitive processing
- The requirement to scroll to view all parts of the item
- Constructed response items requiring longer responses, due to differences in typing speed and ability, typing fatigue, and potentially biases in human markers reading PB and typed responses
- Constructed response items that require students to transcribe calculations from rough paper or calculator to the PC or tablet
- Items with numerical answers that require the "number pad" to input the response
- Items that require use of the on-screen drawing feature for drawing or labelling

(For the original research citations, see Fishbein et al. (2018, p. 7)).

Results showed that among the 'expected invariant' items, there was a general mode effect: digital items were found more difficult than their PB equivalents. Among the 'expected non-

[1] The IRT analysis underpinning TIMSS reporting requires an overlap in items ('trend' items in TIMSS-speak) between TIMSS series. The transition of some test-takers to digital tests therefore required precise knowledge about the psychometric properties of the digital and PB versions of trend items, in order to end up with correctly calibrated TIMSS scores.

invariant' items, there were much larger differences, as anticipated. The solution for eTIMSS was to re-calibrate scores from digital items, so that TIMSS and eTIMSS scores would remain comparable. Although the difference between average PB and digital TIMSS scores was "substantial", the more detailed analyses reached two highly reassuring conclusions. Firstly, the findings showed that the overall maths and science constructs assessed by TIMSS were not affected by the switch from PB to eTIMSS: students found the eTIMSS items more difficult than the PB equivalents, but the same constructs were still being measured (Fishbein et al., 2018, p. 19). Secondly, the analyses showed that mode effects did not have differential effects on students according to socioeconomic status, gender, or digital self-efficacy: "These student characteristics explained a negligible proportion of the variance in achievement score differences between paper TIMSS and eTIMSS" (p. 19).

PISA

PISA assessment, like TIMSS, was originally carried out using PB tests, but from 2015 onwards many countries took the tests in a digital format. Jerrim et al. (2018) analysed PISA 2015 field trial data: pupils in Sweden, Germany and Ireland were randomly assigned to answer either PB or digital versions of the same PISA items, which allowed an estimation of the causal effect of assessment mode on outcomes. The results showed that assessment mode had a "substantial impact upon pupils' performance", with lower scores on digital items. In maths, students taking the digital PISA assessment "scored, on average, around 0.10–0.20 standard deviations lower than their peers who took the paper test" (2018, p. 481). In science, a particularly notable finding was substantial variation between countries in how PB and digital results compared: "the negative effect of taking the test on computer is three times larger in Germany than in Sweden (–0.25 versus –0.07). … To put these figures into context, an effect size of 0.2 is roughly equivalent to 20 PISA test points." (2018, pp. 481-482) The analysis found that the difference between PB- and digital item performance varied by item, as well as by country and by subject. While PISA's scoring methods in 2015 included adjustments intended to account for mode effects in final scores, Jerrim et al. (2018) concluded that the adjustment was insufficient – although better than having no adjustment at all.

Robitzsch et al. (2020) investigated whether the substantial drop in PISA Science scores in Germany in PISA 2015 could be accounted for by the switch from PB to digital assessment, or the adjustments to the scoring models, or both. Like Jerrim et al, Robitzsch et al. (2020, p. 15) concluded that mode effects had materially affected final PISA scores, and that "reported trend estimates between PISA 2012 and 2015 should, therefore, be interpreted with some caution."

1.2.2 Other mode effect research

Experimental studies with adults demonstrate format effects between logically equivalent digital response formats, and modelling supports the interpretation that format affects response strategy (Moon et al., 2020).

To ensure validity of assessment, it is essential to look beyond performance (i.e., marks achieved) on items in different modes, and determine more direct evidence about what response activity is actually elicited by items in their PB and digital formats (Threlfall et al., 2007). Empirical studies looking explicitly at response behaviour tend to be smaller in scale,

but offer rich evidence that complements that from large-scale studies such as the TIMSS and PISA studies noted above.

- With school-age students, findings repeatedly show that the impact of assessment mode on maths item response activity varies across different items (Johnson & Green, 2006; Logan, 2015; Threlfall et al., 2007). Research carried out by Cambridge Assessment International Education with primary age learners has also found that response strategies for some but not all items change in response to assessment mode.
- Differences in response activity arise because the affordances of the on-screen test environment and paper-and-pencil are different. Affordances are defined by Greeno (1998) as "qualities of systems that can support interactions and therefore present possible interactions for an individual to participate in" (p. 9). An example of a possible interaction that is supported differently by PB and digital test environments is an exploratory trial. In their comparison of parallel on-screen and PB maths items, Threlfall et al. (2007) observed that in four of the five items where on-screen performance was higher than PB performance, the item in question involved arranging elements to give a solution. For these items, the ability to carry out 'trial' arrangements was a relative affordance of the on-screen environment over the PB environment, and this type of exploratory activity was often observed among the participants (pp. 341-342).
- Threlfall et al. (2007) emphasised the range of mode effects: while for many items the change from PB to digital format make little difference, for other items "the affordances of the computer profoundly affect how the question is attempted, and therefore what is being assessed" (p. 335). The changes to mathematical activity elicited "in some cases can be sufficiently dramatic to suggest that the nature of the mathematics has changed" (p. 346).
- The way in which mode affects response activity is also not equal for all students. Logan (2015) shows an interaction between mode effects and students' visuo-spatial ability.

1.2.3 Limitations

A limitation of mode effect studies is that many investigate tests consisting of MCQs or short items only, the 'low hanging fruit' in terms of lifting and shifting maths and science items. Considering the list of 'risk factors' for mode effects by Fishbein et al. (2018, p. 7), for example, it seems possible that relatively few GCSE Mathematics questions would be classified as 'expected invariant'. The TIMSS items themselves were all MCQs or short constructed response items, worth a maximum of 2 marks each.

For Drijvers (2019), going beyond the 'easy to assess' questions is a crucial threshold in the digital assessment of mathematics: "can we go beyond straightforward multiple-choice tasks and make students really "do mathematics" in a digital test?" (Drijvers, 2019, p. 43).

2. What are the purposes of the Digital Mocks Service, and how far could they be met for maths and science?

2.1 Purposes

Market research is being carried out by the Digital Mocks Service team with schools involved in the first trials. Initial findings from the GCSE Computer Science trial assessment identified that it had been used by schools for the following purposes:

- To provide an additional assessment opportunity in the school year
- To provide an external assessment opportunity, before live series final exams
- To investigate the practicality and experience of an on-screen mock compared to the usual PB mock
- To give students the experience of on-screen assessment.

Teachers involved with the trial gave very positive feedback on the combination of on-screen tests with external marking from Cambridge examiners. The perceived advantages included:

- Teachers becoming better informed about the assessment standard, and what was credited/not credited with marks
- Gaining a set of marks from the Cambridge examiner that were external to the teacher's own judgements, which:
    - Increased confidence by triangulating teacher marks
    - Were perceived to be bias-free
    - Were perceived to offer back-up to the teacher in their dealings with the rest of the school, particularly leadership, for example in negotiations for resources, or in seeking recognition for departmental achievements.
- Offering convenience for the teacher in assessment set-up
- Significant reduction in teacher time spent on marking
- Significant reduction in teacher time spent on administration (e.g., scanning/posting scripts)

Some teachers in the Computer Science trial commented that not marking their own students work felt "strange" and that it would be useful to have the option to mark the assessments themselves.

Thinking about mock exams more generally, existing knowledge and experience of teaching practices points to the following common purposes:

- To provide a 'practice run' for high-stakes exams
- To inform teachers of a student's likely grade on the live series exam
    - Identifying/confirming which students are not on track for their predicted grades
    - Reassuring students (and parents and teachers), where student is on track
    - Giving certain students a 'reality check' or jolt into action
- To give insights into the student's understanding (e.g., stronger/weaker topics)
    - Informing teacher about where to direct interventions
    - Informing student about where to direct attention in learning/revising

- A possible new purpose: producing evidence that could support teacher assessed grades, which could bring some resilience to the system should we face a situation (like the recent pandemic) where end-of-course examinations cannot take place.

The feedback from the Digital Mocks Service Computer Science trial also confirmed that teachers wanted and expected mock exams to be in the same format as the final exam.

2.2 Possible maths and science assessments in the Digital Mocks Service

There are different solutions to the challenges of lifting and shifting items in maths and science.

2.2.1 Omit items

One option would be to omit items, where the digital 'lifted and shifted' version was considered unlikely to be comparable to the PB version. This could result in the digital test having different construct representation: as documented in previous research, it seems likely that 'expected non-invariant items' would occur more frequently in:

- Extended problem-solving items
- Items involving graphs, complex figures
- Items assessing shape, space and measure
- Items requiring demonstration of skills.

Based on the experience gained from developing progression tests in science, colleagues suggest that this development approach could result in over-representation of the assessment objective on demonstration of knowledge and understanding, and under-representation of the assessment objective on experimental skills and investigations.

2.2.2 Retain altered items

A second option would be to retain items that we expected to not be comparable to their PB equivalents, so long as there were grounds for expecting it to validly assess (some) relevant maths or science construct. For instance, if the shift to the digital platform resulted in an item becoming substantially harder or easier than on paper, or inviting an entirely different response strategy, but the item still assessed a mathematical or scientific construct, it could be retained. Careful communication to schools using the assessment would be necessary, to avoid misinterpretation or the impression that we were unaware of likely differences.

2.2.3 Replace items

Another option would be to try to replace 'difficult to shift' items with combinations of other items that assessed the same content or assessment objectives. The replacement items could be taken from other existing PB exams, if available, or commissioned. This choice would increase the resources required to produce the Digital Mocks Service assessment.

2.2.4 Restrict to those using same learning platform

A further option would be to offer Digital Mocks Service assessments only to classrooms where teaching and learning takes place using the same or an extremely similar digital platform. This would provide significant reassurance on the validity of the assessment, but severely reduce the size of the potential market.

2.3 Which purposes could be met?

None of the options above would offer students a fully authentic ‘practice run’ for their final PB exams, or produce mock grades directly comparable to those from PB exams. However, other valuable purposes of a Digital Mocks Service assessment could still be met, namely providing an insight into student strengths and weaknesses, providing an additional assessment opportunity, and providing an external assessment opportunity – which feedback from the Computer Science trial indicates is valued by teachers.

A clear risk of adding maths and science qualifications to the Digital Mocks Service would be producing assessments that offer an experience that is too different to current high-stakes final assessments in maths and science to be informative, and potentially frustrating or confusing for students, while at the same time lacking the offsetting benefits of auto-marking and assessment innovation that digital assessment could offer in maths and science. The opportunity for external marking might make a Digital Mocks Service product attractive to a school whose maths and/or science teaching and learning already made use of a platform with tools very similar to those in the Digital Mocks Service platform. For schools not currently using such tools in their teaching and learning, the benefits offered by the Digital Mocks Service-style assessments seem much more limited. Our intention is to work with teachers and learners, as well as continue our research, to ensure that any future digital assessments in maths and science support teaching and learning, are informative, easy to use and exploit the benefits that digital technology offers.

Appendix: Big related questions that are out of scope

- To what extent can Maths and Science be assessed on-screen in general?
- What would be the wins, losses, and risks of live assessments for GCSE and A level Maths and Science going on-screen?
- What would be the effects on teaching and learning in these subjects?
    - Unlikely that we would want to remove the handwriting of mathematical and scientific notation altogether, based on what we know about the cognitive mechanisms of language learning (Fernandes & Araujo, 2021; Wiley & Rapp, 2021), and the need to prepare students adequately for progression in maths and science.
- What innovative item types could we design/use to assess the same constructs that we currently assess in PB exams?
    - The opportunities are very significant
- Are there constructs within maths/science that we could assess in a CB format, that we weren't able to assess at all in PB exams?
- Are we equipped to carry out valid auto-marking in maths and science?

Things to consider

- What are the purposes of assessing the knowledge and skills that we assess in maths and science general qualifications?
- What counts as an 'authentic' use of maths and science knowledge? Is the 'authentic' use of mathematics and science always clear? How does preparation for everyday use of mathematics and science (as a citizen) co-exist with preparation for study and careers in STEM?
- What might we be willing to trade, for the innovation that is possible in digital maths and science assessment?

References

Arthur, A., Kapoor, S., & Steedle, J. (2020). *Paper and Online Testing Mode Comparability: A Review of Research from 2010–2020*. ACT Research & Policy Technical Brief.

Aspiranti, K. B., Henze, E. E. C., & Reynolds, J. L. (2020). Comparing Paper and Tablet Modalities of Math Assessment for Multiplication and Addition. *School Psychology Review, 49*(4), 453-465. https://doi.org/10.1080/2372966x.2020.1844548

Backhouse, R., Verhoeven, R., & Weber, O. (1997). Matsad: A System for On-Line Preparation of Mathematical Documents. *Software-Concepts and Tools, 18*, 80-89.

Drijvers, P. (2019). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et évaluation en éducation, 41*(1), 41-66. https://doi.org/10.7202/1055896ar

Fernandes, T., & Araujo, S. (2021). From Hand to Eye With the Devil In-Between: Which Cognitive Mechanisms Underpin the Benefit From Handwriting Training When Learning Visual Graphs? *Frontiers in Psychology, 12*, Article 736507. https://doi.org/10.3389/fpsyg.2021.736507

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*(1). https://doi.org/10.1186/s40536-018-0064-z

Green, C., & Hughes, S. (2022). *Item types and demand: What is the impact on demand of manipulating item types in Computer Science GCSE and IGCSE?* Cambridge University Press & Assessment.

Greeno, J. G. (1998). The situativity of knowing, learning and research. *American Psychologist, 53*(1), 5–26.

Greenwood, L., Cole, U. M., McBride, F. V., Morrison, H., Cowan, P., & Lee, M. (2000). Can the same results be obtained using computer mediated tests as for paper-based tests for National Curriculum assessment? *Proceedings of the International Conference on Mathematics/Science, Education and Technology*, Association for the Advancement of Computing in Education (AACE), 179-184.

Jerrim, J. (2016). PISA 2012: how do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495-518. https://doi.org/10.1080/0969594x.2016.1147420

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476-493. https://doi.org/10.1080/03054985.2018.1430025

Johnson, M., & Green, S. (2006). On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies. *The Journal of Technology, Learning, and Assessment, 4*(5).

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal, 27*(4), 423-441. https://doi.org/10.1007/s13394-015-0143-1

Mills, R., & Hudson, J. (Eds.). (2007). *Mathematical Typesetting: Mathematical and Scientific Typesetting Solutions from Microsoft*. Microsoft Corporation.

Moon, J. A., Sinharay, S., Keehner, M., & Katz, I. R. (2020). Investigating Technology-Enhanced Item Formats Using Cognitive and Item Response Theory Approaches. *International Journal of Testing, 20*(2), 122-145. https://doi.org/10.1080/15305058.2019.1648270

Ofqual. (2020). *Online and on-screen assessment in high stakes, sessional qualifications.* Ofqual/20/6723/1.

Robitzsch, A., Ludtke, O., Goldhammer, F., Kroehne, U., & Koller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. *Frontiers in Psychology, 11*, Article 884. https://doi.org/10.3389/fpsyg.2020.00884

Shepard, L. A. (2008). Commentary on the National Mathematics Advisory Panel Recommendations on Assessment. *Educational Researcher, 37*(9), 602-609. https://doi.org/10.3102/0013189x08328001

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66*(3), 335-348. https://doi.org/10.1007/s10649-006-9078-5

Wiley, R. W., & Rapp, B. (2021). The Effects of Handwriting Experience on Literacy Learning. *Psychological Science, 32*(7), 1086-1103. https://doi.org/10.1177/0956797621993111