CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

A structure for analysing features of digital assessments that may affect the constructs assessed

Research Report

Victoria Crisp
Jo Ireland
14 December 2022

Author contact details:

Victoria Crisp & Jo Ireland
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

vicki.crisp@cambridge.org
jo.ireland@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:

Crisp, V. & Ireland, J. (2022). *A structure for analysing features of digital assessments that may affect the constructs assessed.* Cambridge University Press & Assessment.

Introduction

Digital testing is an area which is receiving increased attention for a variety of reasons. The Covid-19 pandemic and the resulting disruption to exams and teaching brought into sharp focus the need for greater resilience in the assessment system in order to be prepared for times of crisis. Nevertheless, given that such crises should be rare, perhaps other potential benefits should be the key drivers for change. We are now at the point where technology can support us in thinking about how assessment can be improved. Digital assessments might allow certain knowledge and skills to be assessed that are difficult to assess effectively and authentically on paper. For example, digital platforms can potentially record evidence of the process that a learner used to complete a task rather than only showing the final response, thus facilitating assessment of additional skills. Assessing extended written responses digitally may well be more authentic in some subjects if schoolwork tends to be conducted using a word processor. Indeed, assessing extended writing digitally, with the potential to draft and refine, may be more likely to assess the constructs needed in learners' future careers given that the vast majority of writing in professional contexts is likely to be conducted on computer.

Whilst there are a number of possible benefits, digital testing needs diligent preparation, development and evaluation before it can be implemented. Indeed, there may be some knowledge or skill types that might continue to be better assessed on paper, at least until further digital developments emerge. In this report, we draw on existing literature about mode effects to suggest a structure for comparing question design features between paper-based (PBT) and computer-based tests (CBT), where these design features have the potential to influence whether the same constructs are assessed in both modes. This structure is intended to act as a prompt for the types of issues to consider when developing digital tests that use the same questions as, or similar questions to, existing paper-based tests. When moving questions from a paper-based to a digital test we cannot assume that the properties of the question remain the same because changes in the way a question is presented or in response method could affect the constructs that are assessed. If schools and learners are provided with a choice of whether to complete an assessment on paper or on screen with results treated as equivalent, it is important that test developers are aware of any effects on the constructs assessed and on difficulty. Test providers should minimise these effects or be transparent about differences and take measures to address them where appropriate (e.g. equating the tests across modes or stating that there are differences in constructs assessed).[1] If a digital assessment is constructed from questions that originally appeared in an exam paper but will now be used for a different purpose (e.g. a digital formative assessment based on questions from a past paper-based exam), thought should still be given to whether moving the questions onto screen has altered the constructs assessed and, if so, whether there are implications for the intended uses of results. If a digital assessment is designed from scratch and there is no paper-based equivalent or original, then comparability between modes is not a concern. Nonetheless, the structure

[1] See Sireci (2021, p.11) for a discussion of how there could be arguments for relaxing the standard concerns of achieving comparability between modes as long as there is comparability between learners in the inferences made from results from different modes.

proposed here may still provide a tool to think about how features of a task's design might contribute to the constructs assessed and, indeed, whether it is desirable and intended that certain constructs are assessed.

It is hoped that the structure proposed in this report will be useful for identifying design features affecting the constructs assessed in digital assessments, however, it may need to be tailored or expanded for different kinds of digital assessments and/or to address the specific issues that may be pertinent in certain subjects.

Framework for considering features of digital test design that may affect the constructs assessed

As a starting point, we consulted work by Fishbein and colleagues (Fishbein, 2018, Fishbein et al., 2018) comparing TIMMS (the Trends in International Mathematics and Science Study) questions between modes. They identified a set of item types and item features that might contribute to whether items across modes were likely to be equivalent by drawing on existing literature on how learner behaviour may be affected by computer-based testing. Fishbein and colleagues used these types and features as part of a framework for classifying the overall equivalence of items between modes. This qualitative a priori analysis of items was an attempt to predict whether there would be mode effects for certain items.

To support the development of digital assessments, and to encourage reflection on not just whether mode will affect difficulty, but whether mode may affect the constructs assessed, we propose that it is valuable to consider the *ways* in which questions or tasks are different between modes rather than to focus on predicted level of equivalence. Therefore, we returned to the list of item types and item features compiled by Fishbein and colleagues from the literature as the starting point for setting out a framework to support test developers in considering the effects of task features on constructs assessed. Through reference to other literature, including a review by Green (forthcoming), and through viewing examples of digital assessment questions in a number of subject areas, we revised the list from Fishbein and colleagues to provide a structure that we hope is appropriate for use in relation to subject-based qualifications in the UK.

Table 1 shows the proposed structure, along with extracts and points from the relevant literature explaining their origin of each category. The final column provides some possible effects of the features on the constructs assessed. Note that the latter are hypothesised drawing on existing understandings but further research could usefully confirm whether these issues occur in real use of the assessments and the importance of any such effects.

Table 1. Question types and features that could affect the constructs assessed and thus contribute to mode differences

Category	Source(s)	Potential effects on constructs assessed
A. Presentation, format or layout.	• Fishbein (2018, p.63): "Differences in presentation between paper and digital formats (Pommerich, 2004), including items that required significant changes to the formatting to render on a digital interface (Sandene et al., 2005)". • Green (forthcoming): 'technological issues' such as width of lines of text and presentation of items. • Fishbein (2018, p.64): "Difficult response modes, including items requiring students to draw, label, or manipulate features (Sandene et al., 2005; Strain-Seymour et al., 2013)".	• Could partly measure reading skills, computer literacy, familiarity with the platform, attention to detail, and/or working memory.
B. Size/nature of answer space, e.g. answer space structured with space for working out and a final answer line vs one larger answer space on screen.	• Answer spaces within an exam paper affect how much students write in response (Crisp, 2008), presumably through influencing the students' expectations of what is required to do well.	• Could partly measure learners' skills at structuring responses and evaluating how much they need to write.
C. Heavy reading demand.	• Fishbein (2018, p.63): "heavy reading possibly requiring greater cognitive processing (Chen, Cheng, Chang, Zheng, & Huang, 2014; Noyes & Garland, 2008)" (i.e. where a question involves considerable reading this may require greater cognitive processing on screen than on paper).	• Could require greater cognitive processing on screen.
D. Complex graphs or diagrams.	• Fishbein (2018, p.63): "complex graphs or diagrams (Mazzeo & Harvey, 1988)" may increase cognitive processing requirements when presented on screen.	• Could require greater cognitive processing on screen.
E. Scrolling a resource/text extract. **F. Scrolling the question/answer space.**	• Fishbein (2018, p.63): "Scrolling required to view all parts of the item (Bridgeman et al., 2003; Pommerich, 2004; Way et al., 2016), particularly for science items when the student must refer to earlier parts of the item to formulate a response (Pommerich, 2004)". • Green (forthcoming): scrolling when reading – not seeing entire text in one go or all MCQ options, navigation difficulties, scrolling between stem and response, scrolling in response window.	• Could partly measure digital navigation skills and/or working memory.

Category	Source(s)	Potential effects on constructs assessed
G. Long answer. Constructed-response items worth 6 marks or more.	• Fishbein (2018, p.63): "Constructed response items requiring long explanations (Strain-Seymour et al., 2013), due to differences in students' typing abilities (Russell, 1999), typing fatigue that could occur with an on-screen keyboard (Pisacreta, 2013)". • Green (forthcoming): good typing skills provide advantage on essay questions. • Editing and revising responses is easier on screen (Hughes & Greene, 2012) – this is likely to be more important for longer answers. • Planning, structuring, and making an argument tend be better when hand writing rather than word processing an essay and learners may prefer to hand write as it feels more natural (Hughes & Greene, 2012).	• Could partly measure typing and word processing skills – but avoids measuring handwriting skills and makes editing easier.
H. Writing support.	• Tools provided within a digital assessment might provide support not present in a paper-based assessment, e.g.: o Formatting ribbon with bold, italic, indents, etc. – may prompt students to consider how they format and structure a response. o Word count assistance. o Automated spelling and grammar checks.	• May prompt learners to consider formatting and structure. • Automatic word count saves time and interruption, reducing the need for monitoring. • Could reduce measurement of spelling and grammar.
I. Answering requires calculation, numerical answers or mathematical notation.	• Fishbein (2018, p.63): "Constructed response items requiring calculations by hand or with a calculator, which may require students to transcribe calculations from scratch paper to the PC or tablet to receive full credit (Johnson & Green, 2006)" • Fishbein (2018, p.63): "Items with numerical answers requiring the number pad to input the response." This might involve challenges when entering more complex mathematical responses, e.g. fractions or equations.	• Could partly measure skills in transferring information accurately. • Could partly measure typing accuracy, familiarity with digital tools for inputting equations and/or digital literacy.
J. Rubrics, instructions or command words.	• Fishbein (2018, p.64): "some items had directional language that may not apply to the digital format, such as "mark an X" when the digital item required the X's to be typed". • Khan & Shaw (2018) argue that 'medium-independent' instructional language allows students to focus on question content and reduces risks of construct-irrelevant variance in marks.	• Minor changes seem unlikely to affect constructs assessed – as long as learners read the text, meaning is clear, and reading demand is similar. • If reading demand increases or instructions are complex, then reading ability and/or digital literacy could affect marks.

The text that follows provides some exemplification for each category and discusses possible effects on the constructs assessed.

Category A: Presentation, format or layout

Some formatting, layout or presentation features of on-screen questions have the potential to affect reading comprehension and thus the learner's understanding of the question or of a text extract. This, in turn, could lead to partly measuring students' reading skills. Examples of features that could influence ease of reading include:

- Font colour/contrast.
- White space surrounding text (e.g. in text extracts).
- Line spacing.
- Use of bold fonts.
- Paper-based exams often position text extracts on the opposite page to answer spaces or on separate insert sheets, allowing learners to view the extract as they answer. Screen and platform limitations may affect how extracts can be presented in a computer-based test.

Layout and presentation differences relating to response method have some potential to add difficulty and to require additional computer literacy skills, or attention to detail. For example:

- If learners need to click on an option or select from a drop-down list rather than tick a box or write down a letter, this requires on-screen navigation skills and, potentially, familiarity with the functionality of the platform. It also requires care to select the desired option from the drop-down list, particularly if the list of options is longer.
- If learners need to respond by linking boxes with lines rather than writing a letter in a box, this may require familiarity with the item type within the software, or some trial and error to learn how to click and drag to create lines.
- If answering involves working with a diagram or graph, such as drawing on it or manipulating features, this could add difficulty to the task in a computer-based assessment as computer literacy skills will be needed and familiarity with the platform may become important. The usability of the platform will affect how problematic this might be.

There is also potential for limits on the functionality of a platform to more drastically affect how a question is presented and answered. In some cases, it may not be possible to effectively render the intended question in the testing platform. For instance, questions which ask a learner to draw a diagram or complete a graph could be difficult to present and assess on screen depending on platform functionality. Test designers are then faced with how to test certain skills and whether reframing the question (e.g. by replacing a 'complete the graph' question with a multiple-choice question presenting images of completed graphs) effectively assesses the intended construct or results in a threat to validity.

Category B: Size/nature of answer space

The size and any structuring provided within an answer space can indicate to students what is expected of them and, therefore, affect their response behaviours.

In terms of the size of answer spaces, particular on-screen testing platforms sometimes have limited options available for the size of the answer spaces, giving test developers less control than with a paper-based test. This may result in CBT providing spaces that are larger or smaller than those in an equivalent paper-based exam. Whilst learners are by no means

forced to fill a larger space and the response boxes may scroll if learners type more than fits in the visible space, any differences in the size of the spaces presented between modes could lead to differences in the amount that learners tend to write. That said, if mark allocations are presented to learners this could help them gauge an appropriate response length. In CBT, answer spaces may appear as boxes rather than answer lines, which are common in exams and other paper-based tests. Answer lines potentially help give learners a sense of how much they should write. Once a learner starts typing in a response space in on-screen tests, this may also give an impression of how much text fits in the visible space and, thus, how much it might be appropriate to write.

In terms of structuring within answer spaces, paper assessments often provide structure such as response prompts (i.e. labels at the start of an answer line) that can help clarify to students what is expected in their response as well as providing structure. For example, numbered prompts where more than one answer is required, or a prompt on the final line of an answer space for the learner's calculated answer (thus separating this from the preceding working). On-screen testing platforms might or might not be able to support such structuring. Where one version of a test provides structure in the answer space but another does not, there is potential for the unstructured version to require more from learners in terms of their ability to identify the kind of response that is required and to structure the response appropriately themselves.

Category C: Heavy reading demand

Reading on screen, rather than on paper, may require greater cognitive processing and reduce comprehension, particularly for longer texts and where learners are less experienced with reading on screen (see Fishbein's, 2018, discussion of Chen et al., 2014). Whilst reading skills and interpretation of text extracts may be relevant constructs in some subject areas, for other subjects this may increase demands relating to an irrelevant construct.

Category D: Complex graphs or diagrams

Where an assessment task requires interpretation of a complex graph or diagram, this has the potential to increase cognitive processing requirements. There is potential that the cognitive processing requirements are higher when the resource is viewed on screen (Mazzeo & Harvey, 1988, cited in Fishbein, 2018), however, it would be useful for research to explore this issue in contemporary tests given the dramatic change in computer and device use that has occurred over the last few decades since Mazzeo and Harvey's research. Until such evidence emerges, it may be wise for test designers to keep in mind the possibility that complex visual resources may be more demanding to process on screen than on paper.

Category E: Scrolling a resource/text extract

Resources or text extracts may require scrolling to see the whole resource in a CBT. For instance, a question which asks students to compare two of four extracts could be problematic if both extracts cannot be seen on screen at the same time. Another example might be where a diagram and related response options cannot be seen on screen together

and scrolling is required. Where scrolling is needed, there is a risk of assessing digital navigation skills and perhaps also working memory capacity (given that the learner remembering what they read earlier on a page might reduce the need for scrolling). Ensuring that learners are aware of how to adjust browser settings so that they can control whether they can see a complete diagram and/or zoom in to read parts of this may go some way towards reducing the potential for unintended constructs to be assessed.

Category F: Scrolling the question/answer space

Another potential issue could occur when scrolling is needed to see all of the question and the answer space(s). Possible examples include:

- Potentially needing to scroll back up to the question and stimulus when part way through answering.
- Scrolling required when selecting from a drop-down list.
- Needing to scroll within the answer box when writing longer answers that go beyond the size of the box.
- Potential to scroll past a question or answer space partway down a page, particularly where response options or response boxes look quite similar.
- Potential for a learner to think they have reached the end of a page when they have not.
- Retaining separate response spaces on screen to echo a structured answer space provided in an exam paper could lead to scrolling being needed as the question could be out of view when filling in later answer boxes.

As with needing to scroll through a resource, needing to scroll within the question(s) could mean that digital navigation skills or working memory capacity affect scores.

Category G: Long answer

Where longer answers are needed, this may introduce potential for typing and word processing skills to be part of what is measured. However, in paper-based tests there is the potential for handwriting to affect marks, something avoided in an on-screen test. Additionally, when answering questions on screen, it is easier for learners to correct or edit a response, reducing the need for crossing out and indicating changes in other ways (e.g. arrows). Nonetheless, some evidence suggests that handwritten essays tend to include better arguments and to be better structured than essays written using a word processor (Hughes & Greene, 2012).

Category H: Writing support

CBT response spaces can be simple text boxes or answer spaces with word processing features including a formatting ribbon. The latter may provide functionality for font size, bold, italic, underline, text alignment, indents, and so on. These are useful tools, more closely emulating working in a word processor. When answering on paper, candidates can underline text or use indents, for example, so it might seem incomparable if only plain text was possible on screen. However, it is plausible that the presence of a word processing tool might prompt learners to consider how to format and structure a response in a way that a

response space on a question paper does not. If an on-screen test is intended as a direct parallel to a paper-based assessment, then evidence on the potential effects of the formatting ribbon may be useful.

The instructions for some assessment tasks include a suggested word count. In a CBT, word count functionality could be provided. An automatic word count is clearly a useful tool, allowing learners to easily monitor their progress in relation to the suggested word count. In contrast, learners answering the question on paper would need to manually count their words, perhaps either pausing at intervals to do so, or stopping to do so when they think they have a complete answer. Mark schemes may limit the marks awarded if a learner does not write enough, as they will have only partially fulfilled the task. An automatic word count function saves learners the time and interruption of manually counting their words, facilitates 'live' adjustment of their answer approach and could reduce the number of learners who do not write enough. In this way, it could affect the response behaviours of learners taking an online test compared to the paper-based exam.

Spelling and grammar checks are further features that may be available in an on-screen test and can support learners' work where these skills do not need to be tested, but may be best avoided if the aim is for the test to be directly comparable to a paper-based test.

Category I: Answering requires calculation, numerical answers or mathematical notation

Where questions require calculations, learners may be used to conducting their working on paper rather than on a computer screen (even if a calculator is used). Therefore, good practice would suggest that learners should be allowed to use rough paper when completing an on-screen test. If learners conduct working on paper they then need to transfer the answer to screen, and preferably also represent their working on screen so that partial marks can be gained for correct aspects of their working even if their final response is incorrect. This is likely to increase the time taken to answer and adds a potential risk of transfer errors. Needing to provide more complex numerical responses and working, or mathematical notation, may be more difficult on screen depending on the functionality of the software platform[2]. These factors could potentially cause barriers to learners being able to show the extent of their mathematical skills. Skills in transferring information accurately, typing accuracy and familiarity with the digital tools could be measured instead of, or alongside, mathematical skills.

Category J: Rubrics, instructions or command words

The information required in rubrics and instructions for on-screen tests will be slightly different from PBT. There are likely to be instructions to support learners in using the testing platform, such as pointers on how to structure a response, using a scroll bar or on uploading files (if relevant). There may also be instructions about how learners need to respond using the functionality within the on-screen test. For example, an instruction to tick one box in each

[2] See Williamson (2022) for a discussion of moving maths and science items from paper-based to on-screen assessments.

row in a paper-based test might be replaced with an instruction to mark one box in each row with an ‘X’ in a digital test. In addition, instructions relating specifically to PBT (e.g. “Use a black or blue pen”) will be unnecessary in a digital test. It has been argued that instructional language that is independent of mode (i.e. avoiding terms like ‘input’) may be better in terms of allowing learners to focus on the content of the question (Khan & Shaw, 2018).

Minor changes to instructions generally seem unlikely to affect student behaviour or the constructs assessed as long as learners actually read the instructions, the meaning is clear, and the reading demand is minimal. There could be risks if some learners do not read the front page of a digital test because they assume it is ‘standard’ information and are keen to start the test. This could lead to learners missing platform-specific information, such as how to navigate the test or manage options. Furthermore, if additional instructions increase reading demand or are complex, this could lead to reading ability or digital literacy affecting marks.

Conclusion

This report has outlined a proposed structure to support assessment designers and on-screen assessment platform developers in thinking about how features of digital assessments may affect the constructs that will be assessed. It is intended that the categories outlined could also be used to review an on-screen assessment. This may be particularly useful where a computer-based assessment is to act as a direct alternative to a paper-based assessment or has been developed based on questions that were originally written for a paper-based assessment. The set of categories should also be useful in considering computer-based assessments with no paper equivalent. In this case, considering whether the features in each category are present in each assessment task could help to reflect on the constructs that may contribute to marks in practice and hence to avoid construct-irrelevant variance. The proposed structure built on literature review by Fishbein and colleagues (Fishbein, 2018; Fishbein et al., 2018), drew on other insights from the literature and was informed by viewing some examples of on-screen tests in the UK context. The proposed structure may not be comprehensive and may need to be tailored depending on the nature of the assessment and the subject being assessed. For example, if an assessment usually provides a choice of questions, test developers may need to consider how this is best operationalised within the capabilities of the testing platform. There is considerable potential for further research to enhance our understanding of the effects of features of digital assessment questions on what is assessed. For example, Fishbein (2018) cites work by Mazzeo and Harvey in 1988 which suggested that complex graphs or diagrams might require greater cognitive processing when presented on screen, but developments in computer display capabilities since then could have changed this. Additionally, it is of course valuable to pilot new developments with relevant learners. Detailed analysis of learner response behaviours could provide more robust evidence of the constructs beyond the subject domain that are assessed when learners attempt digital assessment tasks involving different kinds of features.

References

Crisp, V. (2008). Improving students' capacity to show their knowledge, understanding and skills in exams by using combined question and answer papers, *Research Papers in Education, 23*(1), 69-84.

Fishbein, B. (2018). *Preserving 20 years of TIMSS trend measurements: Early stages in the transition to the eTIMSS assessment.* Unpublished Doctoral Dissertation, Boston College.

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*(11), 1-23.

Green, C. (2019). *The construct validity of digital assessment items: What the literature says.* Cambridge University Press & Assessment.

Hughes, S. & Greene, V. (2012). Comparing word-processed and handwritten essay composition in examinations. *Paper presented at the annual conference of the Association for Educational Assessment in Europe,* Berlin, November 2012.

Khan, R. & Shaw, S. (2018). To "Click" or to "Choose"? Investigating the language used in on-screen assessment. *Research Matters: A Cambridge Assessment publication*, 26, 2-9.

Sireci, S.G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7-16.

Williamson, J. (2023). *The feasibility of on-screen mocks in maths and science.* Cambridge University Press & Assessment. https://www.cambridgeassessment.org.uk/Images/673946-the-feasibility-of-on-screen-mocks-in-maths-and-science.pdf