To "Click" or to "Choose"? Investigating the language used in on-screen assessment

Rushda Khan and **Stuart Shaw** Cambridge Assessment International Education

Introduction

In this article, we consider the extent to which the language used in on-screen examination questions ought to differ from that of paper-based exam questions. We argue that the assessment language in screen-based questions should be independent of the mode of delivery and should focus on relevant and expected test-taker cognitive processing required by the task, rather than on the format of the response. We contend that *medium-independent* language improves how well a question will measure the knowledge, understanding and/or skills of interest by allowing test-takers to focus on its content rather than on extraneous, potentially contaminating factors such as *technological literacy* and *mode familiarity*. The latter factors may constitute potential sources of construct-irrelevant variance and, therefore, pose a threat to how scores awarded to a performance on a question are both interpreted and used.

'Translated' questions

With the "inexorable" advance of technology (Bennett, 2002, p.1) and its inevitable impact on the format, content and direction of educational assessment (McDonald, 2002), there is a growing desire to translate traditional paper-based tests into ones suitable for on-screen assessment. But what do we mean by a 'translated' test? Do we mean one that mimics its paper-based original and involves the same wording and task on screen and in as close a format as possible to how it appears on paper?

A translated test should, among other things, attempt to maintain the integrity of the specific features of the task or context deemed most likely to have an impact on test performance when replicated on screen. In addition, it must be ensured that the measurement of the intended constructs (skills, knowledge, and understanding) is not undermined by the presence of *unnecessary* technological demands (Chalhoub-Deville, 2003). In an age of digital literacy (Spires & Bartlett, 2012), it is important that the level of technological familiarity is not integral to the construct(s) of interest (Abedi, 2004; Abedi & Lord, 2001, American Educational Research Association; American Psychological Association; National Council on Measurement in Education; & Joint Committee on Standards for Educational and Psychological Testing [U.S.] 2014, p.67)[1]. At the same time, however, the integrity of the constructs must not be threatened by the need to remove construct-irrelevant barriers to test performance (Sireci, 2008, p.84). (See Huff & Sireci, 2001; Li, 2006; Russell, Goldberg, & O'Connor, 2003, for an overview of the mounting concerns about the potential threats to the validity of computerised tests.)

The language of instructions in assessments

It has long been accepted that the information provided in the question input (the material contained in a given test question) and in the question instructions (aspects of the task which provide structure and guidance on successful completion) should be presented to the test-taker in an unambiguous manner (Bachman, 1990; Bachman & Palmer, 1996; Carson, 2000; Crisp, Sweiry, Ahmed & Pollitt, 2008; Shaw & Imam, 2013). One source of test-taker anxiety, according to Madsen (1982), is unclear or ambiguously phrased instructions.

Examination questions will necessarily draw upon a number of factors deemed most likely to have an impact on test performance. Such factors can influence the difficulty of the task and how test-takers will perform. Given the requirement to make certain inferences on the basis of test-taker performance, it is crucial that instructions to test-takers are both transparent and accessible. Well-written instructions make it clear to the test-taker exactly what is being asked of them by the test procedure and task, the nature of their expected response, any time constraints and, in some cases, how their response will be scored. It is especially important to provide clear instructions for more complex and/or less familiar tasks (Bachman, 1990, p.124). Bachman and Palmer (1996, p.121) propose three indispensable guidelines for test question instructions. Instructions should be:

1. sufficiently simple for learners to comprehend;
2. short enough so as not to take up too much of the test administration time; and
3. sufficiently detailed for learners to know exactly what is expected of them.

Distinguishing cognitive from technical command words

In considering the language of instructions used in examination questions, it is helpful to use a natural categorisation which is shown in Bloom's 1956 *Taxonomy of Educational Objectives* (Bloom, Englehart, Furst, Hill & Krathwohl, 1956). Bloom et al. found it necessary and useful to distinguish between command words which relate to the type of question and those which relate to how the test-taker is expected to organise their response. Thus, in the anatomy of a question, there are two types of command words: those which refer to the *cognitive process* (e.g., "identify", "predict", "explain" and "contrast") and those which

[1] Though there are those who contend that computer literacy should be conceptualised as a significant contextual factor interacting with the construct measured in a computer-based language assessment (Jin & Yan, 2017).

refer to *how to respond* – the *technical* language of assessment instructions (e.g., "circle", "tick" or "write"). While cognitive language indicates the kind of content expected in an answer, technical language guides the test-taker on the physical steps by which they should register their response.

Cognitive command words have often been the focus of scrutiny and the meaning of certain command words have been explored in detail (Fisher-Hoch & Hughes, 1996). In contrast, technical commands such as "write" do not seem to have warranted the same discussion: the process of picking a pen up and writing is clear. As a consequence, this category of command word has been relatively neglected. In this example, it may be that the process of writing is so obvious that this aspect appears to be a less meaningful feature of the instruction (though "write" is not the only possible "how to respond" type command word). Consider, for example, when this is appended to its cognitively-laden counterpart:

> *Explain the difference between a metaphor and a simile.*
> *Write your answer here.*

Here, the technical command "Write your answer here" does not appear to add any additional information that would not already be known.

However, when used in place of the cognitive command "Explain", the technical commands seem not to reflect the complexity of cognitive processing required:

> *Write down the difference between a metaphor and a simile.*

We would contend, therefore, that cognitive commands are generally more suitable than technical ones in conveying information to the test-taker. Furthermore, it seems there is little loss to clarity if any, to the sole use of the cognitive instruction. We would expect that using a precise cognitive command could improve the test-taker's score, whereas pointing out the use of a writing instrument would not reasonably be seen to do so. The knowledge that the test-taker has to write an answer in a space would seem to be a necessary undertaking for sitting an examination. Test-takers already have a clear expectation which does not need to be explicitly confirmed.

Replacing a technical instruction with a cognitive one has the positive effect of retaining the most important aspect of the instruction while also not burdening a question with an excess of command words. Two command words in one question could reduce readability and understanding. Shaw and Iman (2013) found that the use of two command words in one question should be avoided as test-takers have a tendency to focus only on the first. It therefore seems prudent to make efforts to limit the number of command words used. Ultimately, clarity of the question should be the most important factor in deciding whether a secondary technical instruction is needed or not.

The quest for medium independence

By *medium independence* we mean whether a feature of the assessment can be said to make sense regardless of the medium of delivery. There appears to be a relationship between the category of command word and whether the question is medium-independent. Cognitive command words are always medium independent while technical command words are not necessarily so. By way of illustration, in the example "Explain the difference between a metaphor and a simile", the cognitive process of explanation is the same both on paper and on screen. It is only the answer input that has changed: writing in the first, versus typing in the second. The question itself is equally clear in both mediums.

However, consider the phrase, "Circle the prime numbers", which contains a technical instruction. It is likely that the mechanism for answering the question on screen would be different. On screen, the requirement might be clicking on a checkbox, for example, rather than drawing a circle with a mouse (which would not be appropriate on screen for usability reasons). Such an instruction is not necessarily medium independent and could be misleading on screen. We may be inclined to change the instruction to "Click on the prime numbers" in order to address the difference. Similarly, for reasons we will explore later in the article, test developers may be inclined to append cognitive instructions with technical ones rather than replace them by including phrases such as "Type in your answer here". This is a technical instruction which would, again, make less sense on paper.

A consequence of translating cognitive commands into technical commands is that we introduce a multitude of instruction styles, across different devices (e.g., laptops, tablets and mobiles), which ask for a demonstration of exactly the same skill. There are a number of immediate concerns with this approach. The first, and most basic, is one of practicality: by suggesting that language ought to be medium-dependent, we accept that assessments are different across modes of delivery. Questions in each mode require checking for their differences in layouts (e.g., word wrapping), and whether appropriate technical terminology has been used. In an international context (in which assessment is delivered through English, for example, which may be a second or even a third language) and with rapidly evolving technology, it is easy to see how this could become problematic. Test developers, particularly those familiar with the language of paper-based assessment, would have to consider an additional challenge of deciding whether "Click on the drop-down" is appropriate language for the given technology, (e.g., touch-screen tablets may require test-takers to "Tap the drop-down").

Each practical concern risks a potential threat to validity. Medium-independent language may be easier to understand when transitioning from paper to screen because test-takers will be familiar with the lexicon, and any new language needs to be used cautiously in order to obviate misunderstanding. Furthermore, the threat to validity from the profligate use of command words is equally applicable to paper and screen: the more words a question employs, the more potential cognitive processing is required, the greater the opportunity for introducing a barrier to clarity of instruction and, therefore, the higher the risk of compromising validity. However, this may not be a hugely problematic issue with very simple sentences (such as "Write down your answer here"). As with paper-based tests, technical command words, when used in place of cognitive ones, could reduce the clarity of content, prompting test-takers to focus on features of the question that have less to do with their cognitive understanding, and more with how they interface with technology.

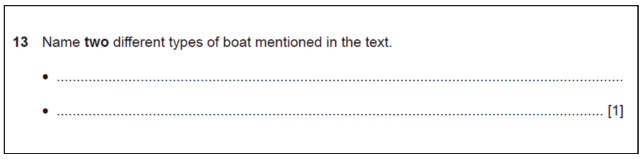
The technological fallacy

Technological fallacy refers to an inherent desire on the part of the test developer to introduce change when responding to the challenges of different delivery formats simply because they are different. The impetus for using medium-dependent language when undertaking direct, word-for-word translations from paper to a digital space appears to be grounded in two imperatives, namely, the *accuracy* and *clarity* manifested in the language of question instructions. Both concepts give rise to at least three scenarios when translating.

SCENARIO 1: There is no perceived requirement for change as the accuracy and clarity in language remains the same across both modes.

In the first example question illustrated in Figure 1[2], there are few perceived challenges in language when translating, as the instruction makes equal sense in both modes. Even though the layout, structure and mode of the response differ, the difference is not deemed to be great enough to modify the instructions. Indeed, the focus is on cognitive processing as the command "Name" does not relate to a particular mode of response.

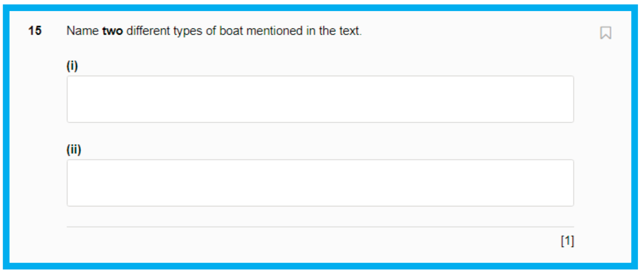
Paper version



13 Name **two** different types of boat mentioned in the text.

- ....
- .... [1]

On-screen version



15 Name **two** different types of boat mentioned in the text.

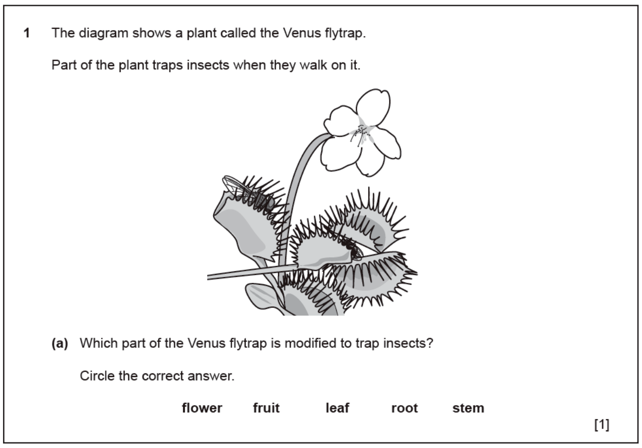
(i)

(ii)

[1]

Figure 1: Example question 1

Paper version



1 The diagram shows a plant called the Venus flytrap.

Part of the plant traps insects when they walk on it.

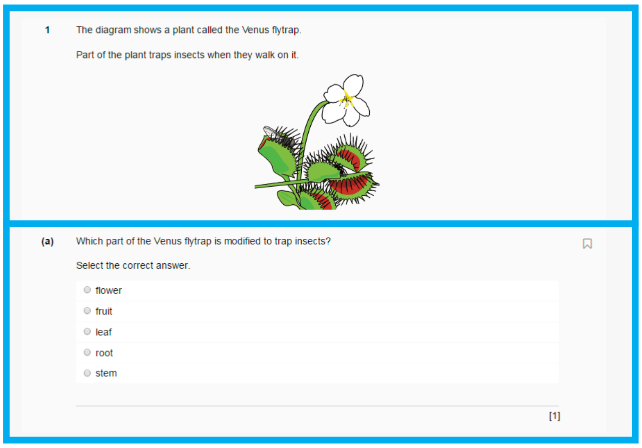
(a) Which part of the Venus flytrap is modified to trap insects?

Circle the correct answer.

flower **fruit** **leaf** **root** **stem**

[1]

On-screen version



1 The diagram shows a plant called the Venus flytrap.

Part of the plant traps insects when they walk on it.

(a) Which part of the Venus flytrap is modified to trap insects?

Select the correct answer.

○ flower
○ fruit
○ leaf
○ root
○ stem

[1]

Figure 2: Example question 2

In this example, a medium-independent approach has been taken. This is appropriate given that the cognitive demand is unchanged and that no mode-specific further instructions are needed.

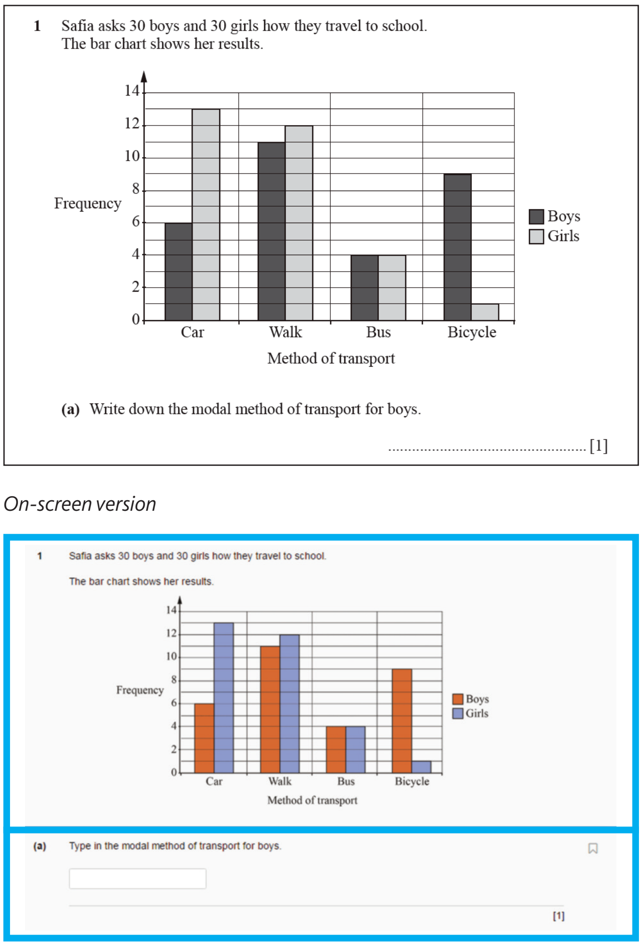
SCENARIO 2: The paper-based language is perceived to lack accuracy and clarity in the new medium but a suggested solution involving medium-dependent language may not be ideal.

In the 'flytrap' example shown in Figure 2, there are a number of differences between the paper version and its on-screen counterpart relating to presentation and response format. In the paper version, the test-taker must "circle" the correct answer while on screen they must "click" on the radio button next to the correct answer. It is important to consider whether this difference necessitates a change in question language. The only difference in the content of the question is the technical instruction "Circle the correct answer", which has been translated to "Select the correct answer". The reason for the change is principally one of accuracy: it would be incorrect to retain the original technical command because circling is not the expected behaviour of a test-taker answering the question on screen.

This is a scenario where the proposed solution is not necessarily the most elegant as it introduces medium dependency. In this scenario, the ideal solution would be to remove medium dependence across both formats and opt for a medium-independent instruction such as "Choose the correct answer". Arguably, the style of response is not intrinsic to the answer so it should not matter whether a test-taker circles the response or indicates their choice in another way. However, for the medium-independent instruction to be appropriate for the paper-based test, the layout of the question would need to be such that there is no risk of it being unclear to a marker which response the test-taker intended to indicate (e.g., if a test-taker ticked in between the options in the paper-based version of the flytrap question, it might not be clear which response was intended). In the current example, a vertical layout of the responses similar to the layout of the on-screen version, could potentially avoid any such issue for the paper-based version, and allow medium-independent instruction to be used for both modes.

In the next example shown in Figure 3, the difference between formats relates to the following instruction: "Write down the modal method of transport for boys". This has been changed to "Type in the modal method of transport for boys" in the on-screen format, presumably in order to enhance accuracy across mediums. This scenario constitutes one in which the imperative to improve accuracy does not, however, require a change in language across formats. "Write" is appropriate language for screen as well as paper. The discipline of digital usability would suggest that we should employ ordinary language for digital tasks as much as possible. For this reason, much (though by no means all) of the digital terminology is based on metaphors from ordinary language. We refer to computer-based 'files',

[2] All examples reproduced for this article are taken from the *Cambridge Lower Secondary* paper and on-screen Progression Tests in development for Mathematics, Science and English. The same product family has been used throughout this article to better identify salient differences. The question numbering may differ between the paper and on-screen versions.



Paper version

1 Safia asks 30 boys and 30 girls how they travel to school.
The bar chart shows her results.

(a) Write down the modal method of transport for boys.

.... [1]

On-screen version

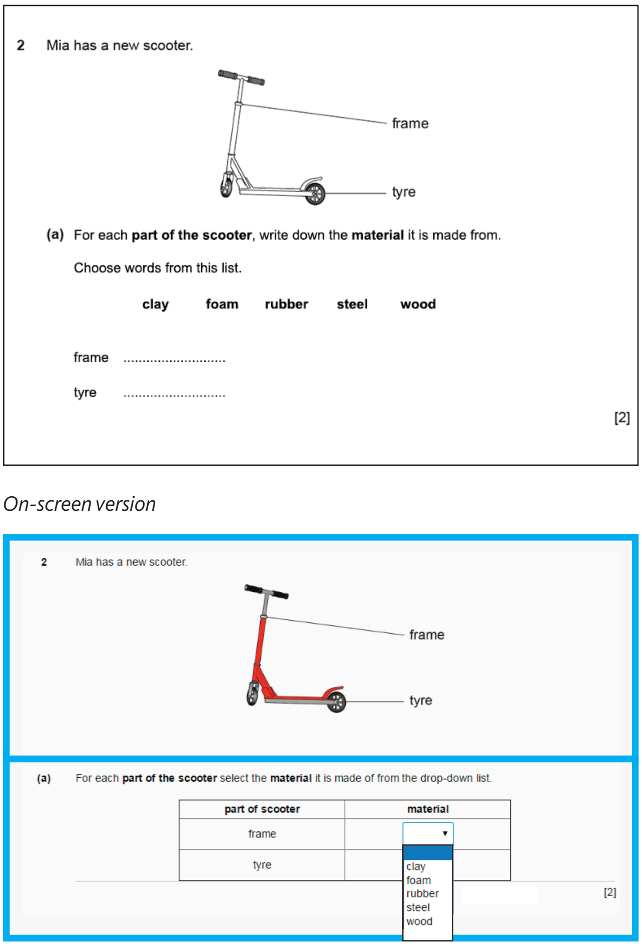
1 Safia asks 30 boys and 30 girls how they travel to school.

The bar chart shows her results.

(a) Type in the modal method of transport for boys.

[1]

Figure 3: Example question 3



Paper version

2 Mia has a new scooter.

(a) For each **part of the scooter**, write down the **material** it is made from.

Choose words from this list.

clay **foam** **rubber** **steel** **wood**

frame ....

tyre ....

[2]

On-screen version

2 Mia has a new scooter.

(a) For each **part of the scooter** select the **material** it is made of from the drop-down list.

part of scooter	material
frame	▾
tyre	

[2]

Figure 4: Example question 4

'folders', 'notepads' and 'recycling bins' which are literally inaccurate but commonplace terminology in a digital landscape. We also use ordinary language for describing digital actions, such as 'posting a comment' and 'writing an email'. This is an illustration of a type of skeuomorphism: the practice of using real world metaphors to create an affordance that increases understanding and familiarity. In this way, counterintuitively, a reliance on the 'correct' technical language can result in an obstacle to comprehension. "Write" would have been clear on screen too. Alternatively, it might have been ideal to remove the technical command completely and the question could have been rephrased as "What is the modal method of transport for boys?" in both formats. This leads to the second reason for introducing technical instruction: the perceived requirement for greater clarity.

The style of the response, in this next example shown in Figure 4, has changed across the paper and screen formats.

On paper, test-takers are expected to pick a word from the list and copy it out, whereas on screen the words are given in drop-down lists for the test-taker to select from. This removes the need for test-takers to transfer the word, slightly changing the nature of the demands of the task. The on-screen version removes the risk that a word might be incorrectly copied over (though incorrect copying should not affect marks given that incorrect spellings of the correct word would be credited. Also, the words are quite different so it should be clear to markers which option was intended). However, the instruction has not only been modified for accuracy but augmented with the technical instruction "from the drop-down list". Presumably, the intention here is to introduce clarity to the question by explaining the correct mode of interaction. We may ask whether this really is necessary. It may be sufficient to use medium-independent language such as "Choose the correct materials" in order to maintain the same level of demand.

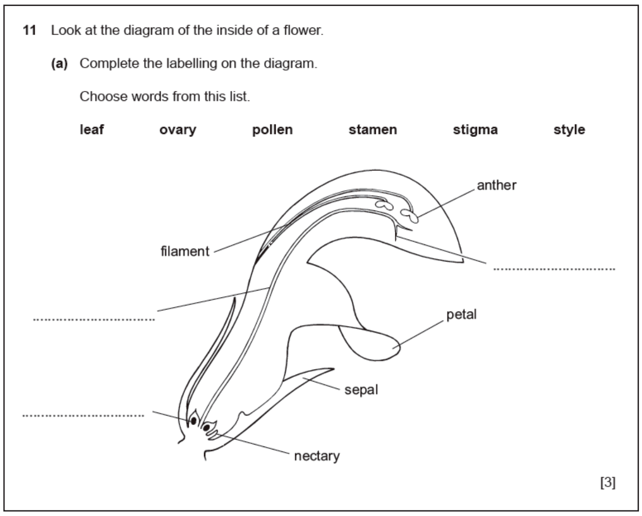
Like the previous example, the next one shown in Figure 5 shows that the style of response has changed between the two modes, but this time, in a more significant way.

On paper, the test-taker must write down the correct words on the label lines, whereas on screen they must drag and drop words from a list into boxes. Again, this slightly adjusts the demands of the task by not requiring re-writing of words and avoids any risk of spelling errors or poor handwriting affecting marks. There is more chance of weaker writing or spelling affecting marks on the paper version of this task (compared to the question in Figure 4) given that several of the options here are somewhat similar (e.g., stamen, stigma and style).

The notion of using medium-dependent language such as "Drag the correct words" may in this case be more compelling. But even in this case, using such words might be counterproductive, as the instruction "Complete the labelling on the diagram using the words below" may be clear enough in both mediums while also being more succinct.

Proponents of the medium-dependent approach may argue that, without showing test-takers the technical mode for answering the question, test-takers may not know what is expected of them. However, this makes an assumption about technical literacy that neglects the notion that the visual cue of a drop-down may be sufficient, and more powerful. A 'clickable' blank box with an arrow

Paper version



11 Look at the diagram of the inside of a flower.

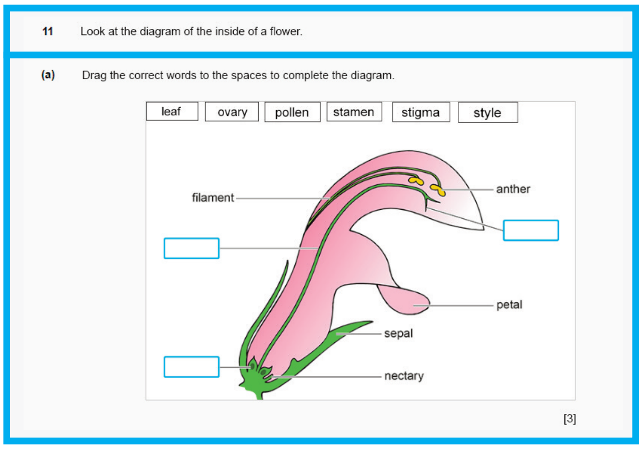
(a) Complete the labelling on the diagram.

Choose words from this list.

leaf **ovary** **pollen** **stamen** **stigma** **style**

[3]

On-screen version



11 Look at the diagram of the inside of a flower.

(a) Drag the correct words to the spaces to complete the diagram.

[3]

Figure 5: Example question 5

that becomes highlighted when hovered over gives a visual instruction in a similar way to lines on a page. We would expect that students encountering lines on a page will be familiar enough with paper to not require an additional instruction of "Write your answer here." If this is the case, it may also be worth considering if a test-taker who is familiar with drop-down boxes from their other digital experiences needs specific instructions.

If basic technical literacy is a requirement of on-screen assessment (as we argue it should be), then it follows that visual cues provide an affordance that is an adequate substitution for a linguistic instruction. It is again part of the technological fallacy that differences need to be consciously accounted for through explicit guidance, rather than acknowledged as differences test-takers can intuitively recognise and account for in their approach to a digital experience.

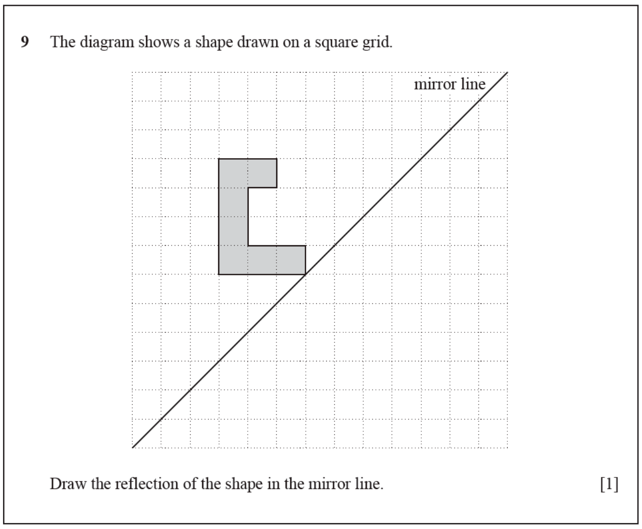
Any requirement to introduce clarity should be in response to a claim that the test-taker may potentially misunderstand the instruction. The most likely way a test-taker may interpret the last example shown in Figure 5, may be that they might try typing directly into the boxes. However, this would result in technical feedback showing the result of the action (e.g., no text appearing), which should prompt them to try dragging instead. We would hope that this would still be a worst-case scenario as test-takers would be, and ought to be, familiar with the notion of 'draggable' words to drop into spaces. To maximise familiarity, design adjustments should first be made to ensure that the interface follows good digital practice. Indeed in this case the draggable words do not look as draggable as they ought to, and it may be clearer to have a 3D shading effect on the words, for example. If pilot use of the testing system or ongoing feedback from schools suggests that some test-takers struggle even after such improvements to the interface, it may be appropriate to encourage schools to use familiarisation activities prior to the test to avoid such misunderstandings.

In summary, if the test-taker's expected technical behaviour to produce a response is unclear, it is preferable to improve the test using a technical solution (such as modifying the design), rather than an assessment solution (such as modifying the instruction). As we will go on to discuss, the reason for a lack of clarity in a technical context is not necessarily due to poor assessment language but poor technical affordances.

SCENARIO 3: The paper-based language is accurate but there is a choice of response methods and it needs to be considered whether this gives sufficient clarity.

So far we have looked at examples where there is a specific way in which a test-taker can answer a question – by filling in a text box or choosing a single word from a drop-down list. But there is another unusual type of question where the test-taker may choose the technical steps with which they will produce their response. Consider the scenario shown in Figure 6.

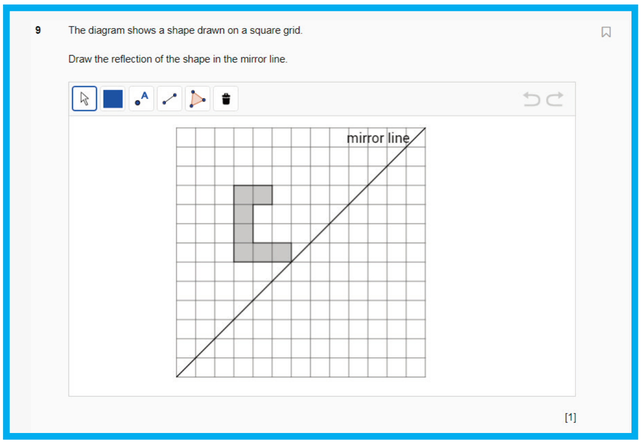
Paper version



9 The diagram shows a shape drawn on a square grid.

Draw the reflection of the shape in the mirror line. [1]

On-screen version



9 The diagram shows a shape drawn on a square grid.

Draw the reflection of the shape in the mirror line.

[1]

Figure 6: Example question 6

In this example, the on-screen version reproduces the paper-based grid but, as is common with on-screen questions, a palette of tools is provided in order for the test-taker to complete the question. There are a number of possible ways in which the test-taker can draw the reflected image, by:

- drawing the points first, then drawing a line to join the points;
- filling in the cells by clicking on each one;
- using the 'polygon' tool to create a shape by clicking on every corner;
- drawing the image freehand; or
- using different combinations of the above.

Any of these approaches may result in a very similar outcome, and it may not even be possible to deduce the tools that the test-taker used based on the completed image. Supplying multiple tools may look peculiar on-screen but, in fact, allowing different approaches arguably mimics a paper approach better than having only one tool: on paper, the test-taker can be as flexible as they wish with the method they use to "draw" their response.

In this case, it would seem particularly undesirable to spell out all the different technical tools in the question itself. A basic attempt may be to append an instruction such as "You may use the line tool, the point tool, the polygon tool or the shading tool to answer the question", but this only announces the suite of tools in the toolbar to the test-taker. More importantly, if the desired approach is to train the test-taker how to use the tools, it would perhaps take up a significant part of the test session to explain each tool in turn, and even then the test-taker would probably like to practise first before committing a response. In general, it is hoped that the tools will be user-friendly and intuitive, and not require much practice.

Thus, introducing multiple potential response techniques makes the argument for including technical commands in the instructions more problematic, as simply signposting them may be unhelpful, and explaining them in detail may be unnecessary and detract from the cognitive demand of the question.

The criterion of technological literacy

As we have argued, attempts can be made to reduce medium-dependent language and it may be helpful to think of a question in each medium in order to achieve this. It also seems that one of the reasons test developers may introduce medium-dependent language is because of an assumption that test-takers need technical guidance in order to understand how to answer digital questions.

Generally speaking, we suggest that test-takers are either 'baseline technically literate' or they are not, and the test developer's approach might affect each category of test-taker differently. Baseline technically literate test-takers are those who are sufficiently technically literate to use technology and are familiar with its conventions. These test-takers are able to navigate to web addresses, recognise and operate scrollbars, open and close windows, and type with confidence. They typically use digital tools on a daily basis for study or leisure purposes. For these test-takers, we may argue that it is unnecessary to provide guidance on how to sit an on-screen test, provided that the quality of the experience is sufficient to follow good digital usability conventions and, therefore, mirror other digital experiences they are already accustomed to. For this reason also, it might be unnecessary to give them instructions on how to use the scrollbar and navigate between questions. If this is acknowledged, then we would argue that these test-takers would also recognise visual cues related to radio buttons, response areas, and drop-down boxes. Any additional explanation of technical facilities should only occur if they are deemed to be more unusual.

For test-takers who are not baseline technically literate, we can assume that they would benefit most from any technical instruction. However, for a test-taker who is not familiar with, or confident to use, computers, it would be unhelpful to introduce specific commands to highlight visual cues on questions. One reason for this is that it may create inconsistency. If, within a question, we need to provide instructions on how to use the computer, then the scope of this must be carefully considered. Otherwise there is no reason why we would only instruct a test-taker to "Use the drop-down list" when we might also need to ask them to "Pick up the mouse", "Hover over the scrollbar", "Press the mouse button", and so on. Additionally, augmenting questions with technical instructions has a direct impact on fairness in a timed exam – time that should be spent on responding to questions is instead spent on learning how to use the technology.

It is likely that there are features of an on-screen test that even baseline literate test-takers would not be accustomed to at all because there is simply not enough of a precedent in their other digital experiences in order to be confident. We have already seen one example of this in Figure 6 where it is unlikely that all students will have encountered the tools they need to use in drawing the response. Another example is if the test-taker needs to write complex mathematical notation using a bespoke toolbar or LaTeX[3] commands. In these cases, an appropriate approach if using on-screen assessments would be to carry out training prior to testing, for example through a familiarisation activity using the tools or copying mathematical notation.

This leads to the conclusion that all digital test-takers need to meet the criterion of technical literacy. Those who take an on-screen test should be baseline technically literate through their prior digital experience. We also need to set boundaries on what counts as familiarisation (and therefore sits outside the test) and what is permitted to take up valuable test question landscape. If the criterion for technical literacy is met 'outside the test' as we would recommend, then test-takers will be able to use the assessment for its intended purposes.

There is one caveat which we have thus far made passing reference to but which ought to be emphasised. It is that the recommended approach increases the burden on the interface developer to ensure that the correct conventions are used in order to maximise usability. For our argument to hold, a drop-down box should look like a drop-down box, and clicking on a single multiple choice option should show one option clicked rather than two. It also requires the interface to provide adequate technical feedback for a test-taker's actions: a state change on hover, the highlighting of a word, and displaying the word in situ when it is clicked. If an interaction is not intuitive due to bad design, it is likely that more familiarisation activities or training will be required, even for baseline technically literate test-takers. This would make prior familiarisation time-consuming and frustrating for test-takers if they have to 'unlearn' good digital practice.

[3] A standard framework for writing mathematical notation on screen.

Quality in a digital landscape has its own guidelines and parameters and a good on-screen assessment should aim to maximise digital usability as well as assessment quality. This can be challenging. Assessment experts are not always digital experts, and vice versa, so there is an inevitable issue in establishing that all the appropriate parameters have been met. We can go some way to address this by acknowledging integrated technical and assessment expertise as essential during the construction of an on-screen test: both types of expertise needs to contribute to its quality assurance and sign off.

The criterion for technological literacy is fundamentally related to the notion of expectation. A suitable test-taker for an on-screen test ought to expect a certain mode of response on each question, either through their general digital familiarity (in which case the test interface is responsible for reflecting their other experiences), or through bespoke familiarisation and training (in the case of a new or unusual digital experience which cannot ordinarily be expected). Unlike the test itself, the familiarisation activities *are* responsible for showing the test-taker how best to use the given technology.

Conclusion

In this article we have explored arguments in support of medium independence in assessment language and recommended a number of key approaches:

- When translating from a paper-based assessment to an on-screen assessment there should not be an automatic translation of cognitive command words to technical ones, or an unconsidered appending of cognitive commands with technical commands. Further research is required, however, to verify such assertions.
- If a command in a translated question does not make sense on screen, it is likely that a technical command has been used on paper. The test developer should consider replacing the technical command with a cognitive command in both cases (or at least using a cognitive command on screen).
- Criteria should be set by awarding bodies for the test-taker to be baseline technically literate before sitting an on-screen test.
- An evaluation of each proposed digital assessment needs to be undertaken to assess whether all features of the digital interface ought to give rise to the correct expectation of a baseline technically literate test-taker. If this is not the case, this could be due to the kind of functionality that:
  a) is in fact *common* but it has been presented in an unusual style; or
  b) is *uncommon* and is bespoke to a new type of on-screen assessment functionality.

If (a) is the case, efforts should be made to follow best practice of digital convention. If this is not possible, or if the likely scenario is (b), then those specific features should form part of familiarisation or preparatory training activities that sit outside the test session.

- On-screen testing sits between assessment conventions and digital conventions and experts in both areas are needed to ensure high-quality assessment.

By way of response to the question raised in the title of this article, a technical instruction like "Click" is an unnecessary modification if the visual cues and technical literacy of the test-taker meet appropriate digital standards. While it is undeniable that a paper-based assessment is different from an on-screen one in a multitude of ways, it is incorrect to say that it is different in *every* way and that the language used necessarily needs to be different. Understanding test-taker expectations and following good practice in technology should allow test-writers to focus on the quality of the assessment, without allowing it to yield to the device upon which it is presented.

Acknowledgements

The authors would like to thank Victoria Crisp, Research Division, for her comments and suggestions on the earlier drafts of this article.

References

Abedi, J. (2004, April). *Differential Item Functioning (DIF) Analyses Based on Language Background Variables*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, USA.

Abedi, J., & Lord, C. (2001). The language factor in mathematics. *Applied Measurement in Education, 14*(3), 219–234.

American Educational Research Association; American Psychological Association; National Council on Measurement in Education; & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bennett, R. E. (2002). Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *The Journal of Technology, Learning, and Assessment, 1*(1), 1–24.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay Co Inc.

Carson, J. G. (2000). Reading and writing for academic purposes, in Pally, M. (Ed.), *Sustained content teaching in academic ESL/EFL*, 19–34. Boston: Houghton Mifflin.

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research, 50*(1), 95–115.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Fisher-Hoch, H., & Hughes, S. (1996, September) *What makes mathematics exam questions difficult?* Paper presented at the annual conference of the British Educational Research Association, Lancaster, UK. Retrieved from http://www.cambridgeassessment.org.uk/Images/109643-what-makes-mathematics-exam-questions-difficult-.pdf

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25.

Jin, Y., & Yan, M. (2017). Computer Literacy and the Construct Validity of a High-Stakes Computer-Based Writing Assessment. *Language Assessment Quarterly, 14*(2), 1–19.

Li, Q. (2006). Equivalence studies of paper-and-pencil based language testing and computer based language testing: A survey. *Foreign Language World*, 114, 73–78.

Madsen, H. S. (1982). Determining the debilitative impact of test anxiety, *Language Learning*, 32, 133–43.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and pencil educational assessments. *Computers & Education, 39*(3), 299–312.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education, 10*(3), 279–293.

Shaw, S. D., & Imam, H. C. (2013). Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations. *Language Assessment Quarterly, 10*(4), 452–475.

Sireci, S. G. (2008). Validity issues in accommodating reading tests. *Jurnal Pendidik dan Pendidikan, Jil. 23*, 81–110.

Spires, H. A., & Bartlett, M. E. (2012). *Digital literacies and learning: Designing a path forward*. Friday Institute White Paper Series 5. Retrieved from https://www.fi.ncsu.edu/wp-content/uploads/2013/05/digital-literacies-and-learning.pdf

Articulation Work: How do senior examiners construct feedback to encourage both examiner alignment and examiner development?

Martin Johnson Research Division

Introduction

This is a study of the marking feedback given to a group of examiners by their Team Leaders (more senior examiners who oversee and monitor the quality of examiner marking in their team). This feedback has an important quality assurance (QA) function but also has a developmental dimension, allowing less senior examiners to gain insights into the thinking of more senior ones. When looked at from this perspective, marking feedback supports a form of examiner professional learning.

This study set out to look at this area of examiner practice in detail. To do this, I captured and analysed a set of feedback interactions involving 30 examiners across three General Certificate of Education Advanced Level (GCE A Level) subjects. For my analysis, I used a mixture of learning theory and sociological theory to explore how the feedback was being used and how it attained its dual goals of examiner monitoring and examiner development.

UK awarding bodies commonly use specialist marking software to distribute digital copies of students' examination scripts to examiners for marking. This allows Team Leaders to monitor the marking quality of the examiners under their supervision throughout the marking period. As part of this monitoring activity, Team Leaders are also required to give examiners feedback on their marking. This monitoring and remediation function is an important component of an awarding body's QA arrangements that ensure that the marking process results in fair and equitable assessment outcomes. An interesting characteristic of recent examiner feedback communication is that it is not generally carried out in face-to-face situations. Feedback is generally given through the software messaging function (i.e., e-feedback), or via telephone communication.

As well as having a crucial QA function, previous work has suggested that feedback can also be conceptualised as having an expansive developmental potential for the less senior examiners (Johnson & Black, 2012). Expansiveness is a concept that describes how some contexts help new participants in a professional community to gain access to the important knowledge and values that then allow them to go on to become more independent participants in an activity (Fuller & Unwin, 2003). I argue, in line with Beighton, Poma, and Leonard (2015); Dennen (2004), and some situated learning theorists, that this concept of expansion has important links to learning, since a development in the understanding of professional practice in an area is synonymous with *learning to be a professional*. This expansiveness includes the type and extent of knowledge transfer, the quality of emotional and practical support for participants, and the appropriate alignment of individual objectives.

Rationale for the study

The acknowledged role that Team Leader feedback has in marking QA processes means that examiner communication is an important area of study. This is particularly the case because of its role in the alignment of Team Leader and examiner thinking which forms the basis of common mark scheme application.

Despite this acknowledged importance, the study of examiner feedback practice is, at present, a relatively under-researched area. This lack of research is the result of a number of specific factors. One factor is that e-feedback practice is still an emerging area of communication, with professional behaviours being inevitably linked to the affordances of the digital marking environments that have recently been adopted across the assessment sector. Another factor links to the challenges of capturing and analysing information that is distributed between individuals across a diverse set of communication channels.

Theory

Learning and communication research suggests a number of potential issues that make the careful study of feedback practice very pertinent.