Writing and reviewing assessment questions on-screen: issues and challenges

Victoria Crisp Research Division and **Stuart Shaw** Cambridge Assessment International Education

Introduction

For assessment contexts where both a paper-based test and an on-screen assessment are available as alternatives, it is still common for the paper-based test to be prepared first. The questions are then transferred into an on-screen testing platform to be as similar as possible to their paper-based counterparts. One challenge with this is that, due to the functionality of the on-screen platform, it may not be possible to transfer certain questions; they either have to be adapted or replaced. This creates additional work and potentially raises comparability issues relating to the equivalence of constructs assessed, level of demand and standards. Moreover, the transfer of items into the relevant on-screen testing platform is often undertaken by personnel other than the subject experts who set the questions. In future, if questions are to be used in on-screen tests, it may be advisable to consider their design from the outset, perhaps with questions drafted by setters directly into the preferred testing platform, with these questions then converted into a paper-based test. This could be accomplished first for easier question types. This could reduce or even remove the need to adapt questions later as the questions will already have been designed to work within any restrictions of the on-screen testing platform. There is also the potential that this strategy could minimise risks for comparability as it is likely to mean that the questions could be more similar between modes (though whether all constructs of interest for a particular assessment context can be assessed through an on-screen testing platform would need to be considered).

If such an approach were taken, drafting assessments directly into an on-screen testing platform might require setters and reviewers to have some different skills and knowledge from those required for paper-based assessments, with potential implications for training and recruitment of personnel. In view of this potential approach, this research explored what is involved in setting and reviewing items specifically for on-screen assessments. The research also provides guidance on the training and/or support that setters would need if writing questions into an on-screen testing platform and that reviewers would need if involved in quality-assuring on-screen test materials.

The research brings corroboration and additional detail to previously existing anecdotal insights.

Method

Assessment context and platform

This research was conducted in the context of Cambridge Assessment International Education Checkpoint[1] tests in Mathematics and Science, with a focus on tests for 14-year-olds (Stage 9). As an example of a well-established on-screen testing platform, Inspera[2] was used in the research. It should be noted that the focus of the research was the knowledge, skills and training that assessors would need in order to set or review test questions using an on-screen testing platform; the research is not an evaluation of Inspera, and most of the themes identified in the current research are likely to apply to other on-screen testing platforms. It should also be noted that the data collection was conducted in April 2018 and there have been updates to the software since then.

Participants

Six participants, three with Science expertise and three with Mathematics expertise, were recruited. All had question setting and reviewing experience in their subject area for qualifications such as Checkpoint, GCSE, IGCSE, O Levels and A Levels. The participants represented a range of years of experience in setting and reviewing. None of the participants had previous experience of writing questions into a platform for on-screen testing. Half of the participants had previous experience of reviewing items in an on-screen testing platform.

The participants were assigned identifiers which are used in reporting some of the findings; Sc1 to Sc3 represent the Science participants and Ma1 to Ma3 represent the Mathematics participants.

Procedure

The research was undertaken in two stages.

Stage 1: Advance familiarisation

The aim of stage 1 was to provide participants with some familiarity with the on-screen testing platform and with the relevant curriculum (if they were not already familiar with it). The participants were provided with:

- login details;
- brief notes on how to create a new set of questions within the platform;
- a link to a website with guidance on using the platform;[3]
- the curriculum for Mathematics or Science, as appropriate;
- three or four example questions targeted at the relevant age group, and;

1. Checkpoint tests provide teachers with feedback on learners' strengths and weaknesses to inform their teaching and measure learners' progress over time. They also enable structured reporting to parents. They are available in English, English as a Second Language, Mathematics, Science and Cambridge Global Perspectives.
2. https://www.inspera.com/
3. https://inspera.atlassian.net/wiki/spaces/KB/pages/57311314/Author

- a form in which to record notes on their initial experiences during the familiarisation.

The participants were asked to explore the authoring area of the platform, try entering the test questions provided, make notes in the form on how they found the software and any challenges encountered, and familiarise (or re-familiarise) themselves with the curriculum.

Stage 2: Research meeting

A research meeting was held with each participant individually. This involved a number of activities:

- *Familiarisation with a workload questionnaire*. Participants read some information about the NASA TLX (Task Load Index; Hart & Staveland, 1988). This is a tool used to evaluate an individual's experience of a task in terms of six dimensions of workload: mental workload, physical demand, temporal demand, performance, effort (fatigue), and frustration.
- *Introductory interview*. Participants were asked about their previous experiences of setting and reviewing assessment questions and to describe key points arising from their familiarisation activities in stage 1.
- *Setting activity (approximately 90 minutes)*. Participants drafted some items into the testing platform as if for a Checkpoint Stage 9 test. They were given a printed copy of the curriculum and free choice of topics to use. The researchers observed the question setting from outside the meeting room using Morae remote observation software (TechSmith, 2011). This allowed the researchers to observe unobtrusively the participant's activities on-screen and a small video feed of the participant and desk area.
- *Completing workload questionnaire*. The participants completed the NASA TLX workload questionnaire in relation to the setting activity. This involved indicating the contribution of each workload dimension to their overall workload in both a relative (i.e., by comparing the contribution of pairs of dimensions) and absolute manner (i.e., by rating the contribution of each dimension).
- *Interview regarding setting activity*. Participants were interviewed regarding the setting activity including their process of drafting questions, differences between setting on-screen items and items for paper-based tests, knowledge and skills needed for setting on-screen, and guidance that setters would need to support them in writing good questions for on-screen tests.
- *Reviewing activity (approximately 75 minutes)*. Participants reviewed a set of questions that had been entered into the testing platform in advance. They were asked to review the quality of the items, whether auto-marking worked appropriately, and whether marking guidance was appropriate. Participants were asked to record feedback on a form (electronically or on paper, according to their preference) to report issues with the questions and suggestions for improvement. Again, Morae was used to remotely observe the participants.
- *Completing workload questionnaire*. Participants completed the NASA TLX workload questionnaire in relation to the reviewing activity.
- *Interview regarding reviewing activity*. Participants were interviewed regarding the reviewing activity. The interview covered their process of reviewing questions, differences between reviewing on-screen versus paper-based items, knowledge and skills needed for reviewing on-screen, and guidance that would be needed to support reviewing on-screen questions.
- *Final interview*. Participants were asked whether they felt that the item setting and reviewing that they had conducted during the research would have been similar or not if they had conducted the same work at home.

Results

As described, there were various different types of data collected in this research (i.e., interviews, observations, completed forms, workload questionnaires). The results have been summarised across these types, given that overlapping themes emerged. The sections that follow focus on the general findings from the advance familiarisation (stage 1) and from the research meetings (stage 2). For the research meetings, the findings are organised by setting and reviewing. Note that the completed forms, observations and interviews revealed various (and overlapping) issues and challenges relating to the specifics of using the platform (e.g., how to enter fractions, how to apply a background image to a hotspot item, how to format tables). These have not been reported in full in each section in order to reduce repetition. Instead, a few examples are given where pertinent, and a full list is presented later to indicate points that would need to be included in training or guidance for setters and reviewers working with on-screen assessments.

Advance familiarisation

Based on their reports, most of the participants were able to start entering questions fairly quickly during their familiarisation work. One of the Mathematics participants struggled to get started, however, and sought help from the researchers. All participants experienced some initial challenges with working out how to use certain features of the platform.

A key theme in reports from the Mathematics participants was the issue of being able to set out questions appropriately. The Mathematics participants tended to think that entering questions was more time consuming for Mathematics than it would be for other subjects, presumably due to features such as tables and mathematical notation. Ma2, for example, observed that some features of Microsoft Word that they found helpful were not present in the on-screen testing software. At the end of the exercise, Ma3 felt confident that they could author a routine calculation question but noted that setting out questions might take up more attention than creating the question content when working in an on-screen testing platform:

> *"Currently … I think if I had to author using this system, my time spent would be more on typesetting and less on creating content."*

Ma3 also expressed concern that:

> *"I don't feel I entered any of the given questions completely. There was some kind of issue on all of them. Those questions were designed for pen and paper so you can't just replicate—they've got to be adapted. More creative ways need to be explored, e.g. if in life we are using a pen and paper then it's natural that they'd write down working out. If in life we are using a computer, then you'd use other tools to solve problems, like spreadsheets, or the internet."*

However, this participant was positive about some aspects of the platform's functionality. Whilst exploring the mathematics function in the editing box, Ma3 noted that the software used was TeX,[4] with which they were unfamiliar. After exploring the functionality of TeX, Ma3 concluded that:

> *"TeX has got the power to do quite a bit more than an equation editor."*

The Science participants also experienced some initial challenges with entering questions. One Science participant (Sc3) felt that they worked more quickly as time progressed through the familiarisation activity. However, they reported that they tended to think of a better way to enter a question (e.g., a different item type) after creating it. This suggests that more experience with the platform may aid setters' abilities to write optimal questions. A concern expressed by all participants related to the frustration of knowing the kind of questions they intended to write but being unable to create them:

> *"I often know what I want to do, however, do not seem able to actually do it."* (Sc3)

There were indications that the participants would have liked to be able to ask someone for guidance on the issues they were experiencing:

> *"At the end, thought I had grasped some of the basics but was left a little frustrated that I could not sort out some relatively minor issues on my own—really needed someone to ask."* (Sc3)

Experiences with the platform during individual research meetings: setting

There were some commonalities in the process setters used when setting questions in the platform as part of the activity, but also some variations:

- Three participants (Ma1, Sc2 & Sc3) started by viewing the curriculum document and either selecting one or more curriculum points to assess or coming up with a question idea at this point. Broadly, this led into consideration of how the question idea or the curriculum point could be assessed using the available item types. Sometimes, the items were then mapped out in notes on paper either with draft wording or just as an idea or rough structure for some parts of a question. Then, participants began inserting items in the platform and drafting question content. Two of these setters usually completed a draft of an item before continuing to the next item, but one setter (Sc2) inserted several items in quick succession and then moved between them adding and editing content.
- Two participants (Ma2 & Ma3) began by selecting an item type to try out. They then viewed the curriculum and selected a curriculum point that they could assess with this item type. One of them then paused whilst they mentally planned the question (Ma3), which was reported to replace their usual and more natural process of developing ideas as they began to write in Word. Participants then inserted and drafted items in the platform. Participants felt that starting by selecting an item type was not an ideal strategy for writing questions but that they were led to this strategy by setting within the platform.
- The remaining participant (Sc1) used a mixture of the two broad strategies described—sometimes beginning by viewing the curriculum but sometimes starting by selecting an item type. Their approach was quite exploratory in terms of investigating different item types and abandoning them if they were not appropriate.

Each participant's process was mapped out as a flow diagram of the steps that they reportedly took to create items during the activity. These are shown in Figure 1. (Note that these do not represent the number of items that participants created.)

Participants experienced a range of challenges during setting. In particular, there were issues relating to the use of 'matching' and 'drag and drop' task types such as locating the tools needed to control features of the questions. Some participants experienced issues with: how to set up auto-marking to give appropriate marks, copying and pasting images, setting up the layout so that there is an initial stimulus before the items, and using TeX. Science participants experienced some problems with placing units after an answer space and with using subscript and superscript in response options (e.g., chemical formulae in multiple choice questions).

Despite some challenges, participants explored and experimented within the platform, for example, trying out different item types, and creating innovative questions (e.g., testing knowledge of the number of electrons in each shell for a particular element using the 'matching' item type). There were positive indications that participants hoped to get better at using the platform with more time. For example, one participant reported that their confidence increased as the session went on.

Differences compared to writing for paper-based tests

During the interviews, participants were asked how setting on-screen items compared to writing for a paper-based test. Participants reported that setting items in the platform made them feel more restricted as their lack of familiarity with the software acted as a barrier. There appeared to be some avoidance of attempting to draft test questions that they were not confident in how to create. Some participants felt that the item types affected their choices around the questions to write. One participant reported writing shorter questions than normal and that they did not develop questions by adding more parts in the same way that they would normally. There seemed to be a tendency to use as many auto-marked questions as possible in order to make use of the advantages of on-screen tests, even though this might change the kinds of questions set compared to their usual setting. It was noted by some that dealing with the software took up part of their thinking and reduced their focus on the subject.

There was some concern about curriculum coverage in that it might be difficult to test some topics or skills within the platform, or that it might simply be more time consuming to set questions on some topics or skills, which would put setters off creating them (especially if payment is per question or per mark rather than per hour). It was also suggested that setters may be less likely to try innovative questions because there is more chance of not being able to create them successfully.

Workload experienced during setting

Participant frustration during setting was a recurring theme across the data collected, a concern that was particularly manifest in the outcomes of the workload questionnaire. As already described, after the setting

4. TeX is a formatting system that is useful for typesetting mathematical formulae, see https://en.wikipedia.org/wiki/TeX (retrieved 26 June 2020). It is provided within Inspera for this purpose.

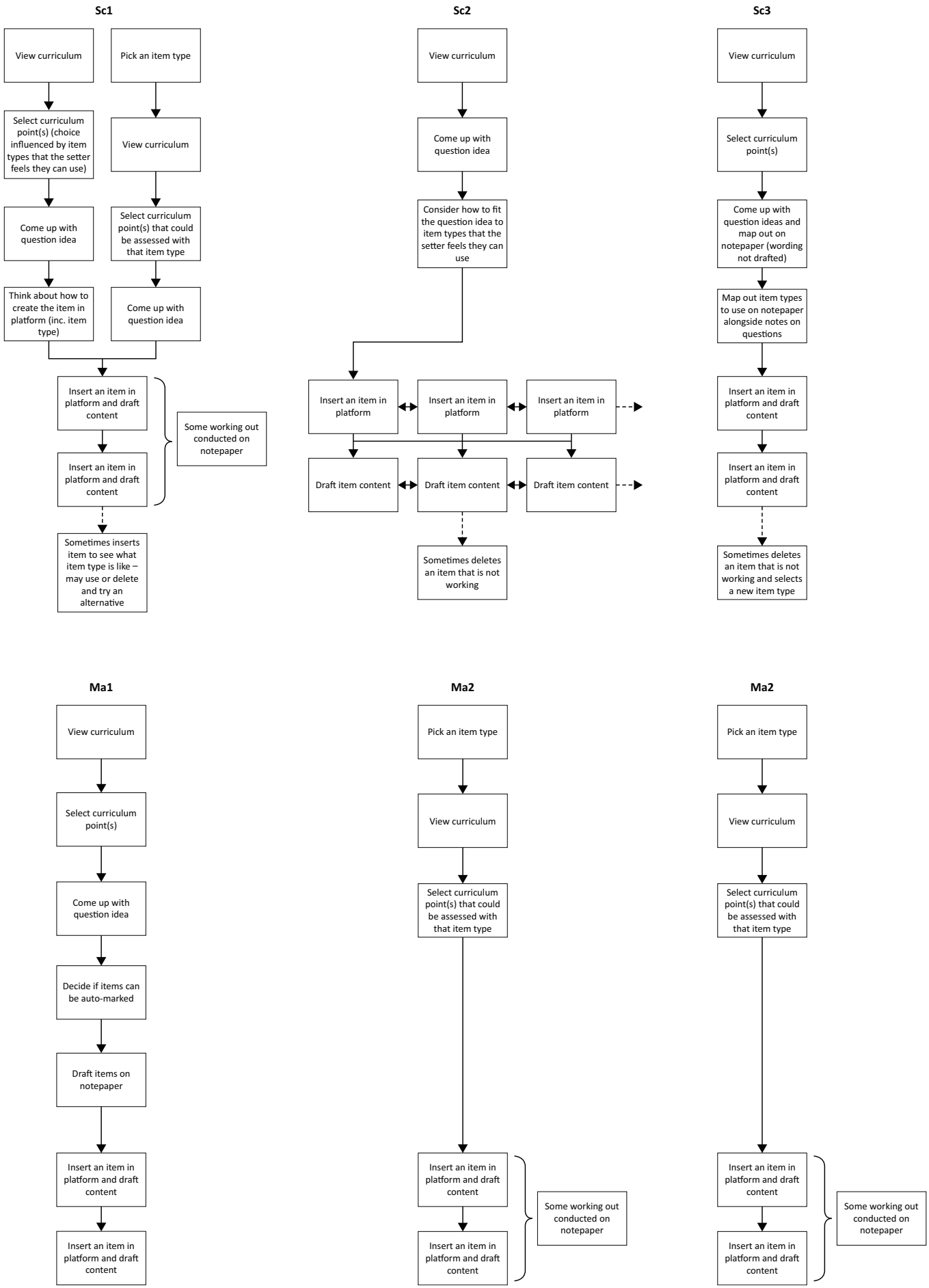

Figure 1: Flow diagrams showing processes during the setting activity.

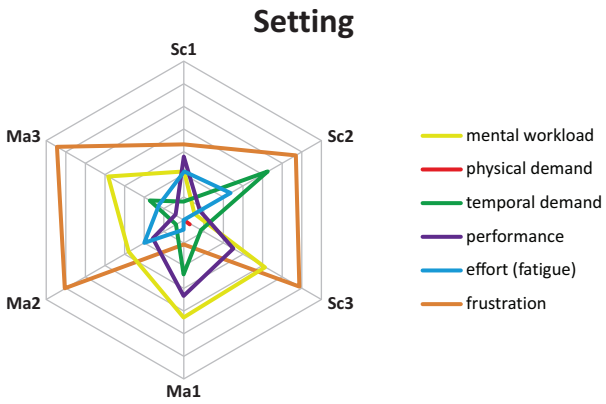

Figure 2: Profiles showing the workload reportedly experienced by participants during setting.[5]

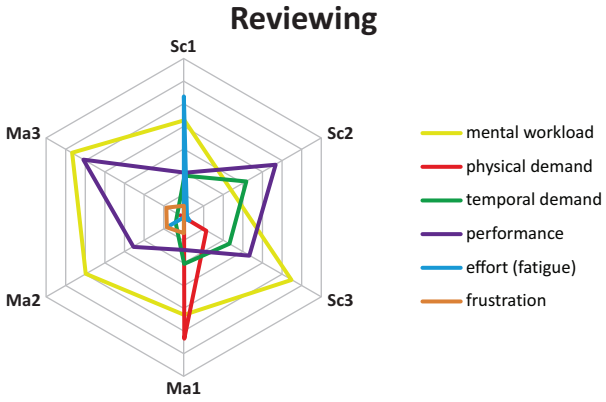

Figure 3: Profiles showing the workload reportedly experienced by participants during reviewing.[5]

activity (and the reviewing activity) participants were asked to report on the workload they experienced in relation to six workload dimensions in both an absolute and relative way.

Workload was analysed according to the usual procedure for NASA TLX workload responses, which for each individual produces workload values for each dimension for setting and for each dimension for reviewing. The values give a measure of the perception of each dimension at that time relative to the other dimensions. There is an underpinning assumption that there is a maximum possible workload and that a high value on one dimension will mean a lower value on one or more other dimensions. In effect, one dimension can 'crowd out' others. It should be noted that as the measures are based on personalised reflections, it is not appropriate to directly compare between individuals. However, common patterns can be considered and comparisons made between activities for one individual.

Figure 2 shows the profiles of the workload experiences reported by the participants during setting. Points further from the centre represent a higher experienced demand for that individual participant relative to the other dimensions. Figure 2 illustrates that for most participants, frustration was the most prominent part of the workload experienced during setting. This workload dimension is defined in terms of how insecure, discouraged, irritated and annoyed the individual feels rather than secure, gratified, content and complacent. These measures are consistent with participants' comments around feeling limited in what they could do due to lack of familiarity with the platform. Mental demand (the amount of mental and/or perceptual activity that is required, for example, thinking, deciding, calculating, remembering, looking, searching) appears to be the second most strongly experienced demand looking at the group of participants as a whole, and was the most strongly experienced workload dimension for one participant. This makes sense given the nature of what is involved in setting test questions.

For one participant, temporal demand was also quite high. This relates to the amount of pressure felt due to the rate at which the task elements occur and to feelings of being rushed. This could suggest that they felt some pressure due to the research situation. With regard to setting, it may be that experiences of frustration have potentially reduced or crowded out other factors such as mental workload and performance for most participants.

5. The axes of Figures 2 and 3 go from 0 (at the centre) to a value of 35 (outer hexagon).

Experiences with the platform during individual research meetings: reviewing

The observations, participant feedback notes about the questions, and interviews provided various insights into the process of reviewing on-screen items. Participants viewed each question checking for various issues, tried out various possible responses, and (usually) viewed the marking guidance. Some participants reported using their usual review process of reading through all the questions first and then conducting a second pass to consider the marking guidance.

When asked about their process of reviewing items, most participants' comments focused on the aspects of items that they were checking. All had a number of elements in mind which appeared to be acting as a checklist as they worked. These included the match to the curriculum, the accuracy of the subject content, spelling and punctuation, appropriateness of command words, formatting and layout, and the functioning of any auto-marking. Some participants considered alternative options for how the questions could be structured, how the stimulus relates to the different parts, and how the stimulus and related items should be set out.

Differences compared to reviewing for paper-based tests

When asked how reviewing questions on-screen compared to reviewing items for a paper-based test, participants generally felt that the process was quite similar in that they were looking for similar things. Some participants reported that they usually reviewed Word documents of paper-based tests and their mark schemes on screen, making the process more similar than if they worked on printed copies. Some differences were mentioned such as considering whether the item type was appropriate, trying out answers for auto-marked items and moving between views in order to check how the item would appear and to check the marking guidance. It was noted that: "What I'm reviewing is different but...what you're trying to do is pretty much the same." One participant felt that reviewing on screen encourages more focus on the marking criteria than on the question.

Workload experienced during reviewing

Figure 3 shows that for reviewing, participants tended to experience a high mental workload (e.g., thinking, deciding, looking) and a fairly high focus on performance (how successful the individual thinks they are in what they have been asked to do and how satisfied they feel with what they accomplish). In other words, participants tended to experience the task as quite challenging in terms of the mental or perceptual activity

involved, and they felt a fair degree of satisfaction in their performance. For one participant, the fatigue of carrying out the reviewing activity was the most strongly contributing factor to their workload, and for another participant, physical demand was most prominent. The reasons for these spikes are unknown. Comparing the two graphs in Figures 2 and 3 (whilst keeping in mind how the values represent relative experiences of the workload dimensions), the clearest difference is that experiences of frustration were generally higher for setting than reviewing. This appears to be offset by greater awareness of mental workload demands and better perceptions of performance during reviewing.

Issues affecting both setting and reviewing

A number of issues affected both setting and reviewing.

Mark allocation

When asked in relation to setting, participants reported that using the platform did not change the mark allocation that they felt was appropriate but that sometimes they did not know how to apply auto-marking so that the mark allocation was correct. Sometimes this led them to adjust the question so that it was worth a mark allocation that they could set up.

Similarly, in relation to reviewing, participants felt that using the platform did not change the mark allocation that was appropriate. Some noted that the software appeared to have affected the mark allocation of some of the items they reviewed, and that the platform might affect what is possible in terms of mark allocation. An interesting comment from one participant was that where a test will be available in both modes, if an auto-marked item gave students more support (e.g., provided response options when the paper-based equivalent item did not) then there could be a case for the on-screen item being worth fewer marks than its paper-based counterpart.

Auto-marking

Participants were asked whether they experienced any challenges relating to auto-marking. Setting up auto-marking was said to be unproblematic for straightforward item types.

One participant reported initial difficulty locating where to put the correct response for auto-marked questions. Another felt that setting up auto-marking was simpler than creating a mark scheme. Participants commented that the need to enter every correct and incorrect spelling that could be accepted was an additional challenge compared to usual. Participants liked that they could try out the auto-marking.

In relation to setting items, participants expressed a need for familiarity with what each item type can facilitate in terms of marking. Specifically, participants mentioned issues with setting up marking for drag and drop items. One participant mentioned that there were challenges with setting up auto-marking where some responses could appear in more than one order (e.g., the products of a chemical reaction in a chemical formula question). Such issues could mean that some questions with short answers might still need to be set up as manually marked questions.

In terms of reviewing auto-marking, most participants felt that checking auto-marking was not problematic and that it was easy to try different responses and see how they were marked. The only challenge mentioned was the need to try out a variety of different options and alternative wordings (e.g., is 'the rope' accepted as a correct answer as well as 'rope'), echoing comments in relation to setting.

Marking guidance

Participants felt that writing marking guidance for manually marked on-screen items would be very similar to writing a mark scheme for a paper-based question. One participant questioned whether a three-column layout, similar to their usual mark schemes, would be needed and how to achieve this (e.g., inserting a table).

When reviewing, some participants initially had difficulty locating the marking guidance but they tended to get better at this with practice. However, they sometimes forgot to check the marking guidance for manually marked questions. Some considered it a challenge or tiring to move between areas. One participant felt that it would be easier to have a mark scheme for the test as a separate document.

Pre-requisite knowledge and skills needed by setters and reviewers

Participant comments suggested that there were very few specific knowledge and skills that were pre-requisites to setting or reviewing on-screen items, beyond the knowledge and skills needed to set or review items for paper-based tests. The pre-requisites that participants mentioned related to general computer skills and confidence, patience and imagination (in order to be able to write questions in a different way).

Training and support materials and what they should cover

In terms of necessary knowledge and skills that could be learnt for the purpose of the role, becoming familiar with various aspects of the platform and its functionality was the main theme. Most participants expressed a preference for face-to-face training, perhaps including some guided learning followed by trying out setting questions with support when needed. One participant felt that some initial independent familiarisation followed by an opportunity to discuss queries would be sufficient. Training could be shorter for reviewers who need to know what the system is capable of but do not need to be able to enter items.

Guidance documents were also mentioned by most participants as potentially helpful with some comments that these should be structured and signposted such that users can easily navigate to find points that they need. There was an interest in having access to some example questions already entered into the platform, and their paper-based counterparts, as a way of seeing how questions can be adapted.

Given the challenges experienced by the participants during the research, setters appear to need training and/or guidance on the following:

- **Setting up the question**
  - The terminology used in the relevant platform;
  - The platform's item types and how to use them (e.g., how to enter fractions or formulae as labels in matching items);
  - How to select an appropriate item type for the question to be asked, especially for common questions (e.g., completing a word equation in chemistry, labelling a diagram);
  - How to test skills relating to handling data and understanding geometrical figures;
  - How to control the layout of questions (e.g., how to display a stimulus alongside the items);

  - How to enter mathematical content (e.g., fractions, formulae);
  - How to create and insert diagrams/graphs/other visuals;
  - How to format tables (e.g., bold in headings, text alignment, width of columns);
  - How to place a unit at the end of the answer line, or text at the start of an answer line (e.g., 'x = _____');
  - How to allow students to enter subscript or superscript into answer spaces (or awareness that this is only possible in certain item types);
  - Which command words should be used (e.g., 'type' or 'write');
  - Whether setters should address formatting or whether this will be checked and refined by a typesetter;
  - How to view stimulus material or earlier items alongside a new item when setting;
  - Whether randomisation of the order of response options should be left on or turned off;
  - Whether 'true'/'false' questions can be used in the test.

- **Setting up marking**
  - Whether a certain proportion of items/marks should be auto-marked;
  - Which kinds of questions common in a subject (e.g., rounding questions in Mathematics, Mathematics questions where the response is a recurring decimal) can be auto-marked and which should be manually marked;
  - How to set up auto-marking to award the appropriate number of marks (e.g., for a matching question where two marks should be given for four correct matches);
  - How to identify all possible correct answers and incorrect spellings that can be accepted (for auto-marked items);
  - Location of marking functions in the platform (e.g., where to set up auto-marking or where to enter marking guidance);
  - How marking guidance should be set out (e.g., in columns);
  - Whether marking guidance should be entered for auto-marked questions.

As mentioned, training for reviewers may not need to be as detailed but some understanding of the platform is needed. Based on insights from the research, training or guidance for reviewers should include:

- the platform's item types and how they can be used;
- navigation in the platform, in particular, locating marking guidance, locating question previews;
- how best to set out screens or move between screens during reviewing such that both auto-marking and marking guidance are checked (e.g., using two internet browser windows so that the item and marking guidance can be viewed side-by-side);
- whether the reviser is responsible for feeding back on formatting issues relating to how the item appears on-screen or whether this will be checked by a typesetter;
- how to identify all possible correct answers and incorrect spellings that can be accepted (for auto-marked items);
- what is possible in terms of awarding marks in auto-marked items.

Discussion

Whilst this research was conducted with one specific on-screen testing software platform and some findings may be specific to that platform, many insights seem likely to reflect issues that setters and reviewers would encounter when beginning to use any on-screen testing platform.

A number of key issues arise from the current research. Firstly, participants varied in how quickly and easily they appeared to be able to get started with entering questions into the platform, suggesting that some setters and reviewers would need more familiarisation and support than others to use an on-screen testing platform.

Participants reported feeling somewhat restricted when setting items due to their lack of familiarity with the platform and that dealing with the software reduced their focus on the subject to be assessed. This aligns with the frustration that appeared to be salient in responses to the workload questionnaire. These issues could reduce with time and increased familiarity with the platform, but possible effects on the questions produced should be considered. There was some evidence that setters may avoid certain item types (potentially reducing curriculum coverage or coverage of assessment objectives), write shorter questions than normal, write less creative questions and, potentially, produce lower quality questions. Any such limitations on the questions produced could mean that a test would not cover the full range of constructs that it was intended to assess (which could in turn affect teaching). The setting process may also be slower, particularly for more complex item types, or if the setter needs to change the item type part way through drafting a question. These points would need careful management in the early days of setters being asked to draft questions into an on-screen testing platform, to ensure the production of a good bank of items assessing the full range of relevant constructs.

Whilst one of the aims of moving towards asking setters to set into an on-screen platform would be to remove the need for questions to be adapted later for on-screen testing, the possible effects of this on the setter's process need to be considered. In the current research, some setters began their process of setting by selecting an item type to use and then working out what content in the curriculum they could test. Whilst research into question setting for paper-based tests indicates that the type of question to be used is often considered alongside the content to be tested when beginning to plan a question (Johnson et al., 2017), the current participants did not feel it was ideal for question setting to be driven by item type. The strategy of starting with an item type might reduce as setters become more familiar with the platform, but if this strategy were to continue to be used at least some of the time, the consequences would need to be considered.

Overall, it seems that it would be possible for setters to create at least some of their questions within an on-screen testing platform such as Inspera. Setting in this way has the potential advantage of avoiding the need for questions that were written for a paper-based test to be substantially versioned or even replaced for an alternative on-screen test, thus saving time and resources, whilst also minimising risks relating to comparability. However, training and guidance would be needed to support setters, and the time and cost investment in this needs to be considered. In addition, there may be initial frustration involved in setting into an on-screen testing platform since it may make setting slower and it may limit the nature of the questions that are produced, especially in the early days. Setter satisfaction should be taken into account as frustration could be an issue.

Another issue to consider is whether it is likely to be productive to ask setters to create all types of questions into the on-screen platform. There is a potential risk, for example, of setters tending to select only item types that are easy to use. Relatedly, whether to ask setters to create questions that would require additional training beyond the basic platform should be considered. For example, within Inspera, an application called GeoGebra can be used to create more sophisticated questions involving graphics, where the aim is to ask students to draw something or add to a diagram. Using GeoGebra would require additional training and it might or might not be worth the setter undertaking this training when balanced against the frequency with which they would use this functionality. An alternative would be to give setters an awareness of what GeoGebra can do, and give them the option to draft some questions in word-processing software with the question later created in the platform by a typesetter with GeoGebra training.

Returning to the finding that participants felt restricted by the platform, it was apparent that sometimes they had ideas for questions that they found they could not implement within the platform. Whilst participants sometimes explored innovative ways to assess concepts, sometimes the restriction they experienced led to compromised decisions about question design that were not satisfactory to participants. This could suggest there is potential for a situation where it is not possible to create questions that tap into certain parts of learning. Over time, if setters can no longer create certain kinds of questions that they would usually write, this could adversely affect content coverage and construct representation. If some individuals are unwilling to make such compromises, they may drop out of involvement in setting. New setters would then be recruited, who might be more accepting of the compromises, thus perpetuating a gradual change in the constructs being assessed. Care would be needed to mitigate risks of this kind in terms of ensuring comparability over time and representation of the constructs contained in the curriculum or syllabus. Asking setters to record question ideas that they could not implement and then working with the software developers to implement appropriate revisions would be one possible way forward.

References

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp.239–250). North Holland Press.

Johnson, M., Constantinou, F., & Crisp, V. (2017). How do question writers compose examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal, 43*(4), 700–719.

TechSmith. (2011). *Usability testing for software and websites*. TechSmith Corp.

A way of using taxonomies to demonstrate that applied qualifications and curricula cover multiple domains of knowledge

Irenka Suto, Jackie Greatorex, Sylvia Vitello Research Division and **Simon Child** Cambridge Assessment Network Division

Introduction

Although they can sound rather grand and esoteric, educational taxonomies are essentially schemes of classification. They are often hierarchical, and provide the terminology that educationalists need to describe and work with different areas of knowledge (so-called 'knowledge domains'). Ever since Bloom and his colleagues created their seminal taxonomy of educational objectives (Bloom et al., 1956; Krathwohl et al., 1964), it has widely been considered good practice to use taxonomies to formulate and review curricula, learning objectives, and associated assessments. Demonstrating sufficient coverage of each of an adequate range of knowledge domains and subdomains is critical for authenticity, for assessment reliability, and for transparency surrounding what students are learning. It is important for regulators, employers and university admissions tutors, as well as the students themselves, to recognise the areas of knowledge, skills and understanding that have been taught and mastered in a particular course.

Bloom et al. (1956) initially created a taxonomy which focused on the cognitive domain. That is, it classified thinking skills as relating to knowledge, comprehension, application, analysis, synthesis, and evaluation. Since Bloom's original work, revisions have been published and alternative taxonomies have been developed to accommodate advances in psychological understanding. Bloom et al. (1956) originally intended to go beyond the cognitive domain, creating a psychomotor taxonomy that focused on physical development. Although they never realised this ambition, some of the more recent taxonomies have done so, covering multiple domains or different single domains. For example, another non-cognitive domain which has been included in some taxonomies is that of interpersonal knowledge, skills and understanding (Hutchins et al., 2013).

Broader domain coverage is important given that many professions and career paths draw upon several different types of knowledge (Bandaranaike & Willison, 2015; Laxmisan et al., 2007; Savic & Kashef, 2013). It seems probable that *general* (sometimes called *academic*) educational taxonomies that cover both cognitive and non-cognitive knowledge domains may also be applicable in applied educational contexts. However, this wider applicability of such taxonomies is relatively underexplored.