# An Investigation into the Impact of Item Format on Computer-based Assessments

**Matt Haigh**
**Cambridge Assessment**

## Abstract

*In this study aspects of item format in a computer-based assessment (CBA) were investigated to see whether they affect the difficulty of the test items. A mixed-methods approach was used: a quantitative strand examined statistical measures of item difficulty and a qualitative strand gathered student perceptions of item difficulty. Parallel forms of a computer-based assessment were administered to 112 students; items in the second parallel form contained modifications of the item format. Difficulty measures of each test item were calculated from the sample and tested for significant differences. In addition, two focus groups of students were used to build up insights into their perceptions of how the item format impacted the difficulty of the items. The study identified that difficulty measures of all the items were very similar in their alternative forms. The focus groups revealed that students were able to readily articulate perceived sources of difficulty.*

## Introduction

In this study assessment validity was considered in relation to CBA. This section outlines the derivation of the research questions starting from an initial consideration of construct validity and construct-irrelevant variance. A more detailed examination of the research literature relating these concepts to item format revealed a lack of evidence regarding the impact of different item formats on assessment outcomes in CBA.

Many authors put validity at the heart of assessment (Popham 2000; Kane 2006) and emphasise the importance of validity in evaluating new forms of assessment:

> *The arguments… regarding traditional and alternative forms of assessment need to give primacy to evolving conceptions of validity if, in the long run, they are to contribute to the fundamental purpose of measurement - the improvement of instruction and learning.*
>
> (Linn et al. 1991, p20)

Arguments have been put forward demonstrating the role of CBA in both enhancing and reducing the validity of test scores. Ridgway & McCusker (2003) highlight benefits of CBA in improving the validity of assessing problem-solving skills, whilst Clarke et al. (2000) identify the detriment to validity from dependence on multiple choice items. Throughout

the history of CBA, there has been discussion regarding the validity aspects of its implementation (Huff & Sireci 2001; Russell et al. 2003).

Educational measurement theory emphasises construct validity in evaluating test outcomes (Messick 1989). Construct validity is defined as "the qualities a test measures, determined by the degree to which certain explanatory concepts or constructs account for performance on the test" (Messick 1989 p16).

Construct validity can be affected by 'Construct-irrelevant variance'; it occurs when the test contains excess variance that is irrelevant to the interpreted construct. For example, a demanding reading stimulus in a science assessment may cause a variance in test scores (related to reading ability) that is irrelevant to the construct being assessed (science).

Some aspects of construct-irrelevant variance have been explored in the CBA literature. A number of studies indicate that students with a good prior knowledge of ICT performed better on computer-based tests (Clariana & Wallace 2002; Russell et al. 2003; Warschauer 2004). Construct-irrelevant variance can be introduced by poor item design (McKenna 2001; Sireci & Zenisky 2006); screen size and resolution (Bridgeman et al. 2003) and the effect of scrolling (Ricketts & Wilks 2002). These studies indicate that aspects of the screen environment or the method of student interaction may be related to sources of construct-irrelevant variance in CBA. Additional research has investigated how the layout of paper-based formats may affect item performance (Crisp & Sweiry 2006) and how screen design affects how website users access information (Helander et al. 1997). However, there is no research on how item format may affect CBA, therefore two research questions were investigated:

- What are the effects of changing the item format[1] on measures of item difficulty with a computer-based test item?
- How do test takers perceive the difficulty of different item formats?

## Method

The research questions required two strands of inquiry. The first question implied a causal relationship between item format and item difficulty, which required a quantitative experimental methodology. The second question focused on individuals' perceptions of item difficulty, which required a qualitative methodology. Therefore a mixed method approach was required to allow the combination of two contrasting methodological approaches (Tashakkori & Teddlie 1998).

The 'Follow-on explanation' design was used (Creswell & Clark 2006); this places the emphasis on an initial phase of quantitative research, with a supplementary second phase of qualitative research to provide supporting information. The initial quantitative phase used a 'post-test/observation only with control group' experimental design (Black 1999). In the second phase a focus group design was used (Watts & Ebbutt 1987).

[1] 'Item format' is the term used in this paper to cover the layout of text, buttons and images on the computer screen, along with the method of interaction used with these screen elements.

## Phase 1 - Quantitative Experimental Design

Two parallel forms of a computer-based test were developed; each test consisted of 15 items based on the GCSE Science curriculum. Five items were identical in both forms of the test to act as a control. The remaining items, shown in Appendix A, were modified in the parallel forms to investigate the effect of the following aspects of item format:

- Presence or absence of colour image.
- Drag and drop categorisation vs. tick-box categorisation.
- Multiple choice single option selection vs. multiple option select.
- Completion by drag and drop vs. drop down selection.
- Matching objects with lines vs. matching objects using a table.
- Static graphic vs. animated graphic.
- Select correct answer vs. drag answer to target.
- Tick-boxes to select statements vs. whole statement selections.
- Visual resources on single page vs. using tabbed panels to move between information.
- Restricted free-text input box vs. unlimited & scrollable free-text input box.

One hundred and twelve students from six secondary schools in England participated in the research; each student was randomly assigned one of the two parallel forms of the test. Two measures of item difficulty were calculated[2] for the items, and evaluated for significant differences across the forms. Note that each of the ten aspects of item design were analysed independently.

## Phase 2 - Qualitative Focus Groups

Two focus groups were held in one of the participating schools. Students were simultaneously presented with the parallel item formats and asked to comment on any differences in difficulty. The subsequent qualitative analysis used word-frequency analysis and open-coding techniques to analyse the focus group transcripts.

# Findings

The science test was taken by 112 students and the seven schools varied in size and school type, but were mainly community comprehensives in urban areas. Background data on measures of student attainment indicated a spread of attainment, although the mean attainment of the sample was higher than the national mean. Control variables relating to student attainment and ICT competence were not significantly different across the two forms of the test. The mean score on the five identical items was not significantly

[2] Using both Classical Test Theory and Item Response Theory paradigms – see Hambleton, R. K., & Jones (1993) for a useful comparison.

different in the two forms of the test, indicating the random assignment had produced well matched samples.

The items in the test were compared using both Classical Test Theory and Item Response Theory measures of difficulty. The focus group transcripts were analysed to evaluate the students' perspectives. These results are presented in the remainder of the findings section.

## Classical Test Theory - Item Facility Analysis

Item facility is the average number of marks achieved by students for an item expressed as a proportion of the maximum mark. A value of 0 indicates a very difficult item; a value of 1 indicates a very easy item. Table 1 shows the facility values for the items in each of the parallel forms of the test along with outcomes of an independent sample t-test to identify significant differences:

**Table 1.** Item Facility Measures

| Item no. | Facility Form 1 | Facility Form 2 | Difference | t-test statistic | Significance |
|---|---|---|---|---|---|
| **1*** | 0.4273 | 0.5088 | -0.08150 | -0.871 | 0.386 |
| **2*** | 0.5364 | 0.5614 | -0.02504 | -0.375 | 0.708 |
| **3*** | 0.7455 | 0.7456 | -0.00016 | -0.003 | 0.998 |
| **4*** | 0.7455 | 0.7456 | -0.00016 | -0.003 | 0.998 |
| **5*** | 0.3136 | 0.3553 | -0.04163 | -0.580 | 0.563 |
| **6** | 0.4545 | 0.4386 | 0.01595 | 0.168 | 0.867 |
| **7** | 0.6473 | 0.6316 | 0.01569 | 0.302 | 0.763 |
| **8** | 0.2727 | 0.2982 | -0.02552 | -0.296 | 0.767 |
| **9** | 0.8182 | 0.8684 | -0.05024 | -0.803 | 0.424 |
| **10** | 0.6045 | 0.6053 | -0.00072 | -0.011 | 0.992 |
| **11** | 0.5233 | 0.4111 | 0.11214 | 1.254 | 0.213 |
| **12** | 0.6727 | 0.7895 | -0.11675 | -1.394 | 0.213 |
| **13** | 0.2273 | 0.2807 | -0.05343 | -0.674 | 0.502 |
| **14** | 0.5000 | 0.6081 | -0.10811 | -1.109 | 0.271 |
| **15** | 0.3063 | 0.4274 | -0.12104 | -1.416 | 0.161 |

*indicates common item

Although differences in difficulty were observed in the parallel forms of each item, the t-test indicates that these were not statistically significant. This suggests that the modifications to item format had very little effect on item-facility in any of the cases.

## Item Response Theory - Difficulty Analysis

In Item Response Theory, the difficulty of an item is established using a common scale, called a 'latent trait', onto which items can be placed in terms of their difficulty and students can be placed in terms of their ability. The model assumes that the difference

between a student's ability measure (on the scale) and an item's difficulty measure (on the same scale) is related to the probability of the student correctly answering the item. The higher the student ability measure is, relative to the item difficulty measure, the greater the probability of the student getting it correct.

The difficulty values for the modified items in each of the parallel forms of the test are shown in Table 2. (In this Item Response Theory analysis the common items were assumed to have identical difficulty so the output for these items is omitted)

**Table 2.** Item Difficulty Measures

| | **Form 1** | | **Form 2** | |
|---|---|---|---|---|
| **Item No** | **Difficulty** | **Standard Error** | **Difficulty** | **Standard Error** |
| **6** | 0.327 | 0.289 | 0.518 | 0.281 |
| **7** | -0.365 | 0.138 | -0.752 | 0.119 |
| **8** | 1.445 | 0.341 | 1.236 | 0.309 |
| **9** | -1.103 | 0.224 | -1.397 | 0.259 |
| **10** | -0.245 | 0.116 | -0.288 | 0.113 |
| **11** | 0.281 | 0.201 | 0.609 | 0.198 |
| **12** | -0.725 | 0.306 | -1.321 | 0.342 |
| **13** | 0.894 | 0.188 | 0.781 | 0.172 |
| **14** | 0.384 | 0.203 | 0.123 | 0.212 |
| **15** | 1.071 | 0.189 | 0.547 | 0.165 |

Differences in item difficulty are evident; however, the accompanying standard error values indicate that these are not statistically significant. This reinforces the interpretation associated with the item facility analysis findings.

## Qualitative Analysis

In the focus groups students discussed the alternative formats of each item. Their comments were analysed to identify whether there were any common themes associated with item difficulty. The comments were divided into two groups:

- Comments associated with reasons for item formats being easier
- Comments associated with reasons for item formats being more difficult.

A word-frequency analysis was carried out and presented using visualisation software[3] whereby a word's frequency dictates the font size in the diagram. The outcomes are shown in Figure 1 and Figure 2.

[3] Wordle : www.wordle.net (accessed 12 April 2011)


**Figure 1.** 'Easy' word cloud

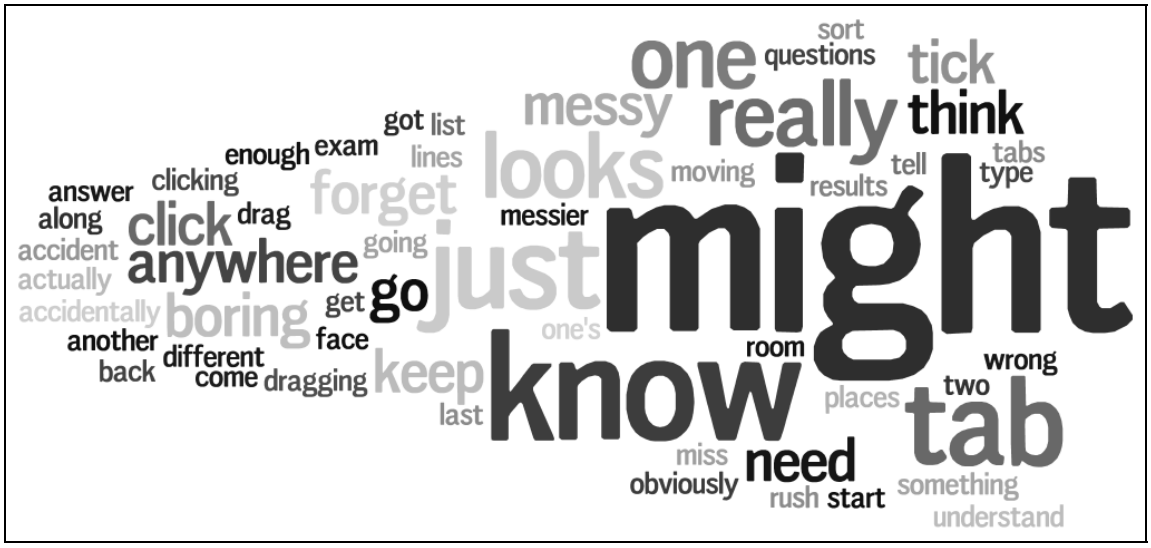



**Figure 2.** 'Difficult' word cloud

The word clouds illustrate that the terminology used around describing items as easy and difficult is very different. In the 'easy' cloud, visual words are prominent ('see', 'look' and 'attention'), as are aspects of cognition ('think' and 'answer'). In the 'difficult' cloud the aspect of uncertainty dominates with a very prominent 'might'.

A shortcoming of simple word-frequency analysis is that it ignores the underlying semantics of the data. To address this, the same groups of comments ('easy' and 'difficult') were open coded and analysed to identify common themes. The emergent codes and their frequencies are given in Table 3.

**Table 3.** Code Frequencies

| **Easy** | | **Difficult** | |
|---|---|---|---|
| **Code** | **Frequency** | **Code** | **Frequency** |
| see all | 6 | lack of clarity | 8 |
| clarity | 4 | misplace response | 8 |
| simple | 4 | unsure of action | 5 |
| eye-catching | 3 | boring | 2 |

| **Easy** | | **Difficult** | |
|---|---|---|---|
| **Code** | **Frequency** | **Code** | **Frequency** |
| help/hint | 3 | distracting | 1 |
| makes you think | 3 | insufficient space | 1 |
| stops waffling | 3 | makes you think | 1 |
| visualisation | 3 | several clicks | 1 |
| focus attention | 2 | | |
| helps review answers | 2 | | |
| common-sense | 1 | | |
| easy to follow | 1 | | |
| easy to understand | 1 | | |
| fewer-clicks | 1 | | |
| fun | 1 | | |
| more opportunity | 1 | | |
| no-clicks | 1 | | |
| one-click | 1 | | |

Some interesting contrasts appear in this table. Firstly, 'see all' and 'clarity' are ranked at the top of 'easy' group and 'lack of clarity' is ranked at the top of the 'difficult' group. One curious result is the presence of 'makes you think' in both categories; on the one hand, if an item 'makes you think' then it is easier because the student is more likely to make a considered response, on the other hand, if an item 'makes you think' then it must be difficult.

In examining the focus group transcripts in detail, one of the most notable themes was the individual nature of the factors identified; in some cases the students' perceptions were similar but many items in the test did not attract a consensus on aspects of difficulty.

## Discussion & Implications

The outcomes indicate that there was little effect on quantitative measures of item difficulty when the item format was changed. The qualitative phase indicated that students were able to identify numerous potential sources of difficulty in different item formats, and that 'lack of clarity' was the most prominent feature of items classified as difficult. The focus groups showed that test takers can have meaningful insights into difficulty associated with item format and could differentiate it from difficulty associated with the subject knowledge demands of an item.

When looking across the two phases of the research, it is possible that the instruments in the quantitative phase were too crude to pick up the nuances given by students in the second phase regarding the difficulty of items. Alternatively, it is possible that that the students' perceptions of difficulty in the focus group setting were less significant in the test setting.

It could be argued that the lack of significance observed in the quantitative data means that the item format makes little difference to the difficulty of the item. However, this would be a simplistic implication given the limitations of the context and the qualitative data that points to subtly different perceptions of difficulty at an individual level. There is already evidence that poor item-design has an impact on the validity of test scores (Huff & Sireci 2001); therefore, it is important to establish which changes to item format would constitute 'poor item-design' and which would make little difference to the validity of the test.

If large-scale, high-stakes examinations move from paper-based formats to CBA, it is imperative that the effects of item format are well understood to ensure fairness to the students undertaking the assessments. It is important that those with responsibility for administering CBA in classroom contexts are aware of the range of student perceptions of difficulty when presented with different item formats. In particular, item design may not have a noticeable effect on the average score in a class of students, but individual students may respond very differently to a specific item design.

The qualitative outcomes indicate that students are aware of how differences in item-format may affect those taking the test and, thus, can have a useful contribution to make in the development of computer-based test items.

## Limitations

In the qualitative phase, it is unlikely that the students were thinking specifically about 'construct-irrelevant variance' when discussing the items in the focus groups and they may have some difficulty understanding the concepts of 'difficulty' or 'easiness'. However, the qualitative data enables a link to be made between the student experience of the test items and the concept of item difficulty.

The following indicates areas where generalisations would be more difficult to make, and also suggests areas for further research activity to allow for wider understanding:

- Subject: This study used items assessing the GCSE Science curriculum; it would be useful to understand if other subject areas raised similar findings and issues.
- Item-types: This study modified ten aspects of item format; however, the effects may be tied to particular item formats, so a wider study of additional factors could be undertaken.
- Sample: This study was constrained to 15 year-old secondary school students in England, and although a reasonably broad sample of these was achieved, the effects may be different in other student populations.
- The impact of ICT competence: This has not been explicitly explored in this study and there could be more scope for identifying its role in the perceived differences in item difficulty.

## Conclusion

The aim of this study was to investigate how aspects of item format in a computer-based assessment affect the difficulty of the test items. All ten aspects of item design that were considered in the study showed no significant difference in measures of item difficulty when administered in parallel forms to the cohort of 112 students.

A supplementary aim was to gain some insight from test users into the nature of any observed effects. The qualitative phase of the research provided some useful insights into the test users' perceptions of difficulty.

The mixed-methods approach provided two perspectives on the concept of item-difficulty. Firstly, the educational measurement perspective showed that modifications to item format made very little difference to the measures of item difficulty. Secondly, the student-centred perspective showed that students have very individual perceptions of item difficulty, although some common traits relating 'difficulty' to 'uncertainty' and 'easiness' to 'clarity' were evident.

The implications are that the measures of item difficulty appear to be relatively unaffected by the item format presented to the student. However, this study also suggests that test-takers themselves have a meaningful contribution to make to item development.

## References

Black, T., 1999. Doing quantitative research in the social sciences: An integrated approach to research design, London: Sage.

Bridgeman, B., Lennon, M.L. & Jackenthal, A., 2003. Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. *Applied Measurement in Education*, 16(3), pp.191-205.

Clariana, R. & Wallace, P., 2002. Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), pp.593-602.

Clarke, M.M. et al., 2000. Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, 32(2), pp.159-181.

Creswell, J. & Clark, V., 2006. *Designing and conducting mixed methods research*, London: Sage.

Crisp, V. & Sweiry, E., 2006. Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, 48(2), pp.138-154.

Hambleton, R. K., & Jones, R.W., 1993. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), pp.38-47.

Helander, M., Landauer, T.K. & Prabhu, P.V., 1997. *Handbook of human-computer interaction* 2nd ed., Amsterdam: Elsevier.

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

Kane, M., 2006. Validation. In R. L. Brennan, ed. *Educational Measurement*. Westport, CT: Praeger, pp. 17-56.

Linn, R.L., Baker, E.L. & Dunbar, S.B., 1991. Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher*, 20(8), pp.15-21.

McKenna, C., 2001. Introducing computers into the assessment process: what is the impact upon academic practice. In *Proceedings of the Higher Education Close Up Conference, Lancaster University, Lancaster, UK*.

Messick, S., 1989. Validity. In R. L. Linn, ed. *Educational Measurement*. New York: Macmillan.

Popham, W.J., 2000. *Modern Educational Measurement* 3rd ed., Boston: Allyn & Bacon.

Ricketts, C. & Wilks, S.J., 2002. Improving Student Performance Through Computer-based Assessment: Insights from recent research. *Assessment & Evaluation in Higher Education*, 27(5), pp.475-479.

Ridgway, J. & McCusker, S., 2003. Using computers to assess new educational goals. *Assessment in Education*, 10(3), pp.309-328.

RUMM Laboratory, 2006. RUMM2020 [computer software]. Available at: http://www.rummlab.com/.

Russell, M., Goldberg, A. & O Connor, K., 2003. Computer-based Testing and Validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, 10(3), pp.279-293.

Sireci, S.G. & Zenisky, A.L., 2006. Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation. In S. M. Downing & T. M. Haladyna, eds. *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 329-347.

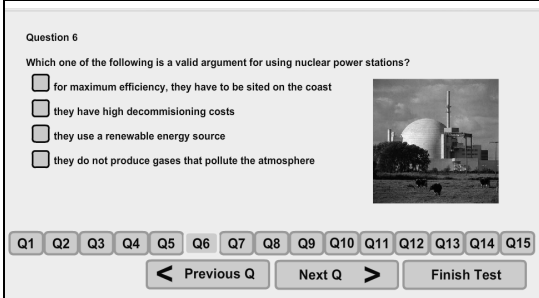
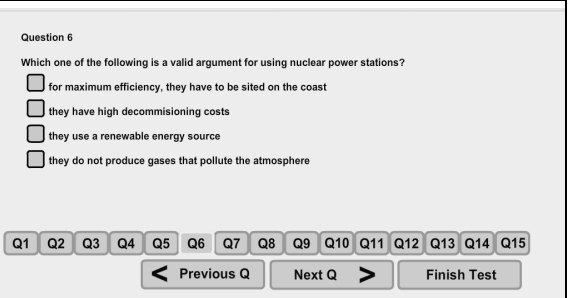
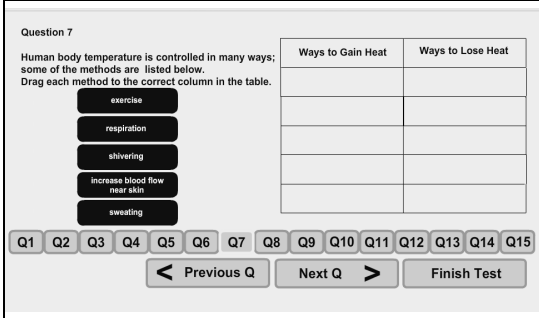
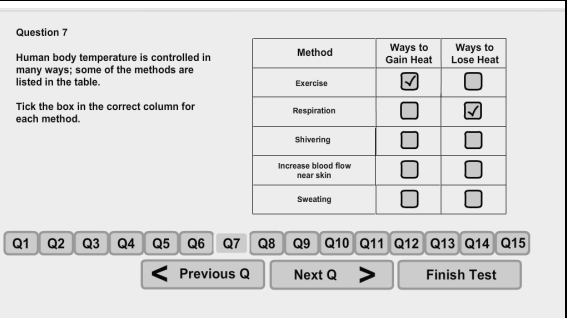
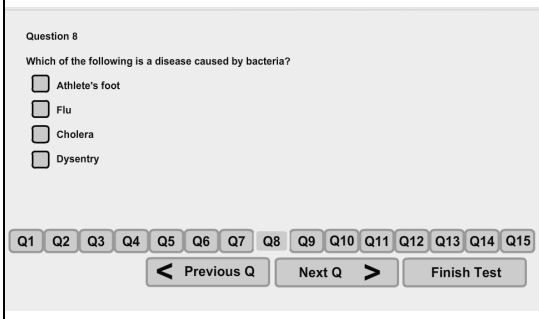
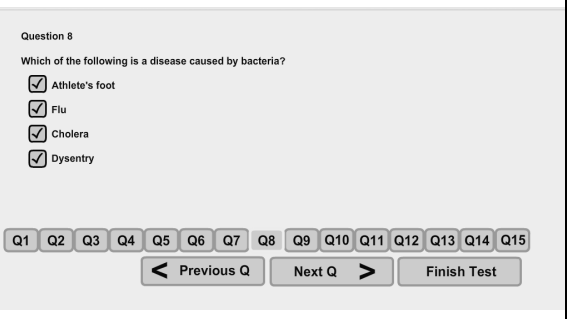
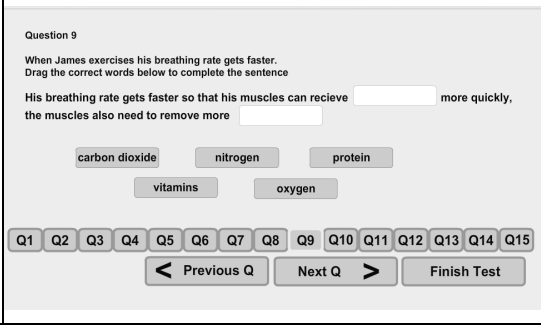
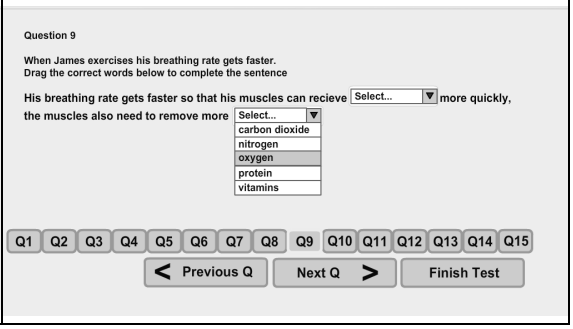
Tashakkori, A. & Teddlie, C., 1998. *Mixed methodology: Combining qualitative and quantitative approaches*, Thousand Oaks CA: Sage.

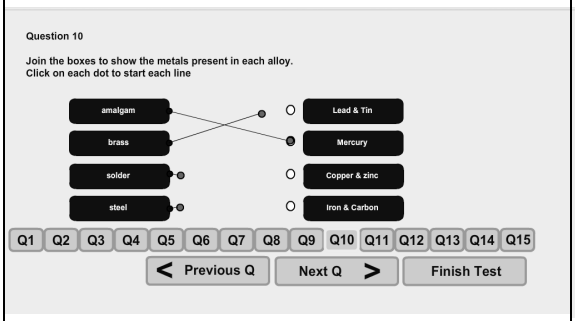
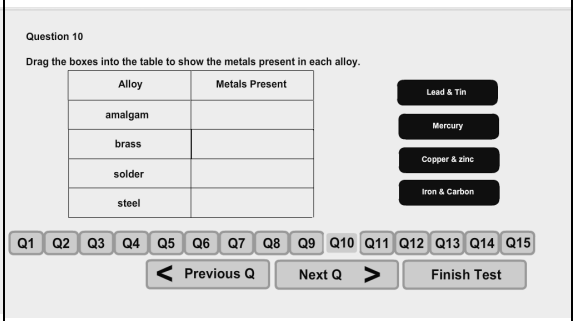
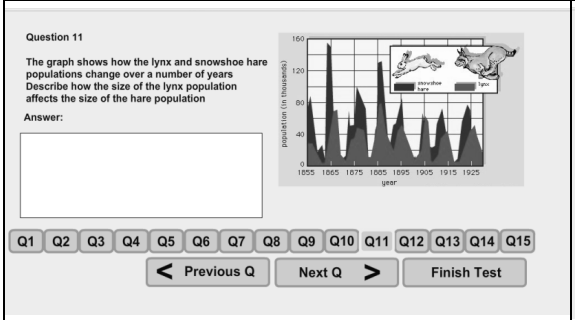
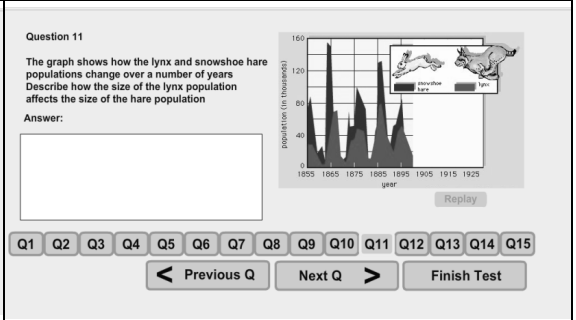
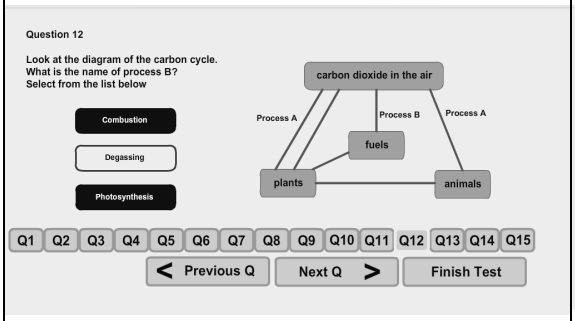
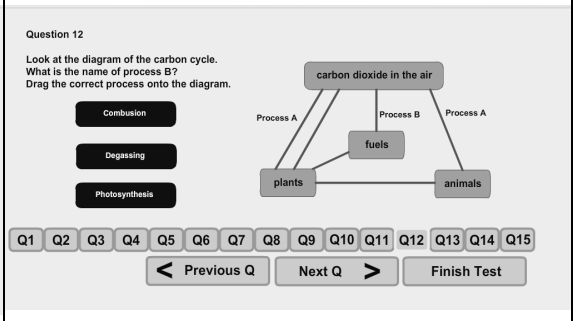
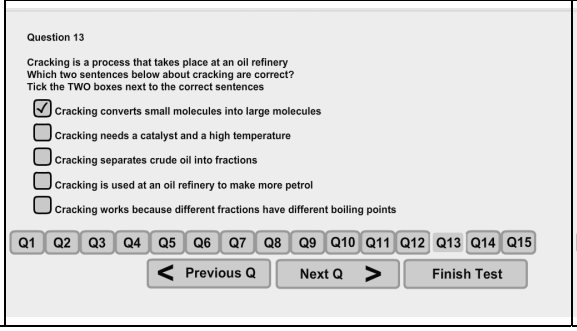
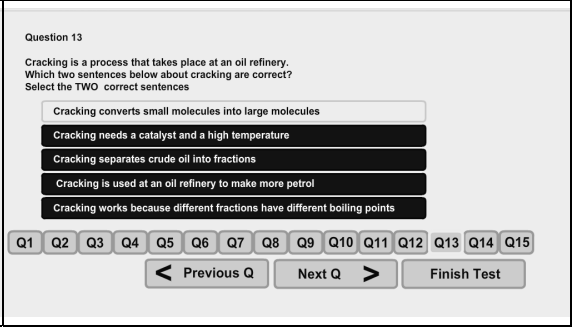
Warschauer, M., 2004. *Technology and social inclusion: rethinking the digital divide*, Cambridge, MA: MIT Press.

Watts, M. & Ebbutt, D., 1987. More than the sum of the parts: research methods in group interviewing. *British Educational Research Journal*, 13(1), pp.25-34.

# Appendix A: Modified Test Items

Common test items (Q1-Q5) are identical on both forms and therefore omitted.

| | **Form 1** | **Form 2** |
|---|---|---|
| **Q6** | Supporting colour image provided | No supporting image |
| |  |  |
| **Q7** | Drag and drop categorisation | Tick box categorisation |
| |  |  |
| **Q8** | Multiple choice, single selection only | Multiple choice, multiple selections enabled |
| |  |  |
| **Q9** | Drag and drop to fill in the blanks | Drop-down selection to fill in the blanks |
| |  |  |

| | | |
|---|---|---|
| **Q10** | Matching options with lines | Matching objects, drag and drop into a table |
| |  |  |
| **Q11** | Static graphic | Animated graphic with replay option |
| |  |  |
| **Q12** | Select option response | Drag answer to target response |
| |  |  |
| **Q13** | Tick-box multiple choice | Select statement multiple choice |
| |  |  |

| | | |
|---|---|---|
| **Q14** | All required data presented on one screen | Required data accessed by tabbed panels |
| **Q15** | Limited text box input | Unlimited text box input with scroll-bar |