Cambridge International Examinations (2011). http://www.cie.org.uk/qualifications/academic/uppersec/preu/subjects/subject/preusubject/?assdef_id=1018

Confederation of British Industry (2007). *Shaping up for the future: The business vision for education and skills*. www.cbi.org.uk/bookshop Accessed 06/04/11.

Department for Education and Skills (2005). *The 14–19 Education and Skills White Paper*. London: HMSO.

Dweck, C.S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Taylor and Francis/Psychology Press.

Ertl, H., Stanley, J., Huddleston, P, Stasz, C., Laczik, A. & Hayward, G. (2009). *Reviewing Diploma Development: Evaluation of the Design of the Diploma Qualifications*. Research Report No. DCSF-RW080. Oxford: Oxford University Press.

Joint Council for Qualifications (2011). News release for A, AS and AEA results, summer 2011. http://www.jcq.org.uk/attachments/published/1584/JCQ%20Results%20Press%20Notice%2018%20August%202011.pdf Published 18/08/11.

International Baccalaureate Organisation (2004). *Diploma Programme assessment: Principles and practice*. Cardiff, Wales: International Baccalaureate Organisation.

International Baccalaureate Organisation. (2007). *Diploma programme extended essay - guide (for examinations in 2009)*. Cardiff, Wales: International Baccalaureate Organisation.

International Baccalaureate Organisation (2011). http://www.ibo.org/diploma/curriculum/core/ Accessed 03/06/11.

Mehta, S., Suto, I., Elliott, G., & Rushton, N. (2011). *Independent Research at A level: Students' and Teachers' Experiences*. Presented at the annual conference of the British Educational Research Association, 6-9 September, 2011.

OCR (2011). *Level 3 Extended Project: Centre handbook/specification*.

Ofqual (2011). http://www.ofqual.gov.uk/qualifications-assessments Accessed 03/10/11.

Tomlinson, M. (2004). *14-19 Curriculum and Qualifications Reform: Final Report of the Working Group on 14-19 Reform*.

Wellcome Trust, (2008). http://www.wellcome.ac.uk/News/2008/News/WTX052388.htm Accessed 04/10/11.

An investigation into the impact of screen design on computer-based assessments

Matt Haigh Research Division

Introduction

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

> *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, p.20)

Arguments have also been put forward demonstrating the role of computer-based assessment (CBA) in both enhancing and reducing the validity of test scores. Ridgway and 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 and Sireci, 2001; Russell, Goldberg, and O'Connor, 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, p.16).

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 and Wallace, 2002; Russell *et al*., 2003; Warschauer, 2004). Construct-irrelevant variance can be introduced by poor item design (McKenna, 2001; Sireci and Zenisky, 2006); screen size and resolution (Bridgeman, Lennon, and Jackenthal, 2003); and the effect of scrolling (Ricketts and 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 and Sweiry, 2006) and how screen design affects how website users access information (Helander, Landauer, and Prabhu, 1997). However, there is no research on how item format[1] may affect performance by students on a computer-based test. This article reports on part of a study that investigated the impact of item format on the difficulty of test items. The following research question was investigated:

> *What are the effects of changing the item format on measures of item difficulty of a computer-based test item?*

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

Method

The research question implied a causal relationship between item format and item difficulty, which required a quantitative experimental methodology. Within this paradigm, a 'post-test/observation only with control group' experimental design (Black, 1999) was used.

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 the Appendix, 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.

Students from seven secondary schools in England participated in the research; each student was randomly assigned one of the two parallel forms of the test. For each item, two measures of item difficulty were calculated[2]. Each measure was then evaluated for significant differences between the alternate forms of each item. Note that each of the ten aspects of item design were analysed independently.

Findings

Sample

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. Table 1 shows the background data relating to the sample.

Table 1: Background variables relating to the sample

Measure	*National Mean*	*Form 1 Mean (n=55)*	*Form 2 Mean (n=57)*	*Form 1 Standard Deviation*	*Form 2 Standard Deviation*
National Test Score[3]	18.49	20.16	19.73	2.49	3.21
Predicted GCSE Score[4] Science	4.64	6.42	6.29	1.24	1.31
Total GCSE Point Score[4]	43.20	57.20	55.84	22.10	23.71
ICT competence[5]	n/a	2.22	2.21	0.83	0.98
Score on 5 common items	n/a	7.24	7.67	3.18	3.12

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

3 Sum of KS2 English level, KS2 Maths level, KS3 English level, KS3 Maths level

4 GCSE score: Grade A*=8 points, Grade A=7 points, B=6, C=5, D=4, E=3, F=2, G=1

5 Self-reported on scale 1= 'Not very good with ICT' to 5 = 'Very competent with ICT'

Measures of student attainment indicated a spread of attainment within the sample, 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 different in the two forms of the test, indicating the random assignment had produced well-matched samples.

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 2 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 2: Item Facility Measures

Item no.	*Facility Form 1*	*Facility Form 2*	*Difference*	*t-test statistic*	*Significance*
1*	0.43	0.51	-0.08	-0.871	0.386
2*	0.54	0.56	-0.02	-0.375	0.708
3*	0.75	0.75	0.00	-0.003	0.998
4*	0.75	0.75	0.00	-0.003	0.998
5*	0.31	0.36	0.05	-0.580	0.563
6	0.45	0.44	0.01	0.168	0.867
7	0.65	0.63	0.02	0.302	0.763
8	0.27	0.30	-0.03	-0.296	0.767
9	0.82	0.87	-0.05	-0.803	0.424
10	0.60	0.61	-0.01	-0.011	0.992
11	0.52	0.41	0.11	1.254	0.213
12	0.67	0.79	-0.12	-1.394	0.213
13	0.23	0.28	-0.05	-0.674	0.502
14	0.50	0.61	-0.11	-1.109	0.271
15	0.31	0.43	-0.12	-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. A visual representation of the data is shown by the scatter plot in Figure 1. The numerical labels on the data points correspond to each of the *modified* items (1 = Question 6, 2 = Question 7, 3 = Question 8 etc.) and the diagonal dotted line represents item forms of equal difficulty. The scatter plot shows all items were close to the line of equal difficulty in their alternative forms.

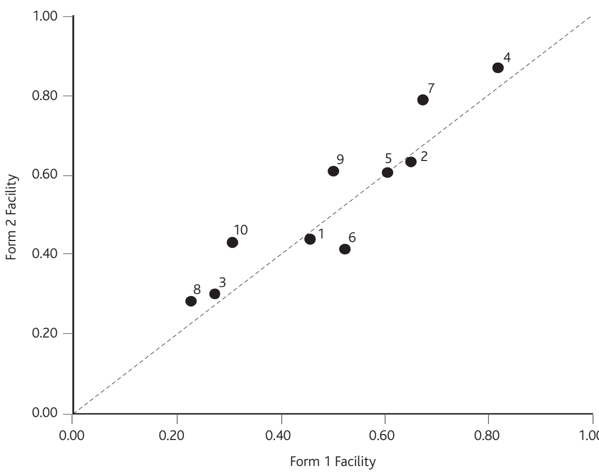


Figure 1: Scatter plot of item facility on each form (Classical Test Theory)

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 3. (In this Item Response Theory analysis the common items were assumed to have identical difficulty so the output for these items is omitted.)

Table 3: Item Difficulty Measures

Item No	*Form 1*		*Form 2*	
	Difficulty	*Standard Error*	*Difficulty*	*Standard Error*
6	0.33	0.29	0.52	0.28
7	-0.37	0.14	-0.75	0.12
8	1.45	0.34	1.24	0.31
9	-1.10	0.22	-1.40	0.26
10	-0.25	0.12	-0.29	0.11
11	0.28	0.20	0.61	0.20
12	-0.73	0.31	-1.32	0.34
13	0.89	0.19	0.78	0.17
14	0.38	0.20	0.12	0.21
15	1.07	0.19	0.55	0.17

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. A visual representation of the data is shown by the scatter plot in Figure 2. The numerical labels on the data points correspond to each of the *modified* items (1 = Question 6, 2 = Question 7, 3 = Question 8 etc.) and the diagonal dotted line represents item forms of equal difficulty. The scatter plot shows all items were close to the line of equal difficulty in their alternative forms.

Note that the two approaches to measuring difficulty produce similar outcomes, items labelled 3, 8 and 10 emerge as the most difficult, and

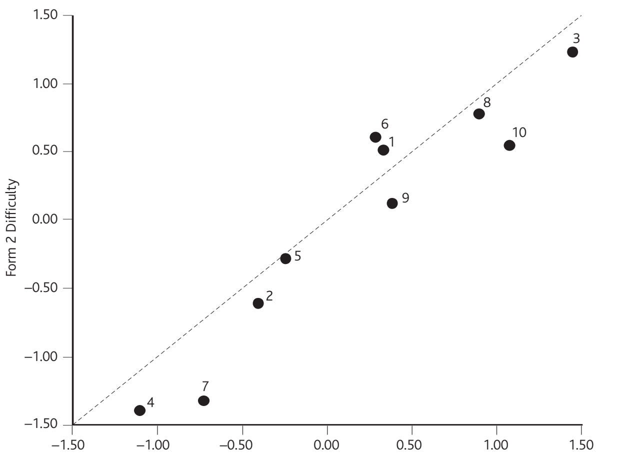


Figure 2: Scatter plot of item difficulty for each form (Item Response Theory)

items labelled 4, 7 and 2 emerge as the easiest, although the rank order of items varies slightly according to which assessment model is selected.

The impact of ICT competence

The research literature indicated that pupil performance on computer-based tests may be influenced by their competence with ICT (Clariana and Wallace, 2002; Russell *et al*., 2003). Therefore in this study students were asked about their level of competence with ICT. Students were grouped into two subgroups according to their responses: 'ICT High' (n= 43) for those reporting that they were very competent with ICT and 'ICT Low' (n=69) for those reporting that they were less competent with ICT. The difficulty of test items was analysed using these subgroups. Table 4 shows the outcomes of this analysis along with tests for statistical significance.

Table 4: Differences in item difficulty by ICT competence group

ICT Group	*Item no.*	*Facility Form 1 High: n=20 Low: n=35*	*Facility Form 2 High: n=23 Low: n=34*	*Difference*	*t-test statistic*	*Signifi-cance*	*Signifi-cant at 5%?*
ICT High	6	0.46	0.50	-0.04	-0.351	0.726	No
	7	0.68	0.65	0.03	0.542	0.590	No
	8	0.31	0.35	-0.04	-0.336	0.738	No
	9	0.84	0.88	-0.04	-0.532	0.597	No
	10	0.62	0.56	0.06	0.712	0.479	No
	11	0.60	0.35	0.25	2.242	0.029	Yes
	12	0.80	0.82	-0.02	-0.247	0.806	No
	13	0.23	0.28	-0.05	-0.503	0.616	No
	14	0.45	0.63	-0.18	-1.521	0.135	No
	15	0.28	0.44	-0.16	-1.499	0.141	No
ICT Low	6	0.45	0.35	0.10	0.669	0.507	No
	7	0.59	0.61	-0.02	-0.197	0.845	No
	8	0.20	0.22	-0.02	-0.137	0.892	No
	9	0.78	0.85	-0.07	-0.641	0.526	No
	10	0.58	0.67	-0.09	-0.916	0.365	No
	11	0.35	0.50	-0.15	-1.071	0.294	No
	12	0.45	0.74	-0.29	-1.977	0.055	No
	13	0.23	0.28	-0.05	-0.439	0.663	No
	14	0.64	0.57	0.07	0.363	0.720	No
	15	0.36	0.40	-0.04	-0.247	0.807	No

The analysis shows that there were generally no significant differences between the test forms for the two ICT competence groups. One exception was Question 11 which appeared to be significantly different in difficulty for the High ICT group only, with form 2 being significantly more difficult than form 1. This item contained a static artwork in form 1, whereas form 2 was modified to include an animated artwork. The observed difference in difficulty is a curious result that goes against the hypothesis that those with better ICT ability are able to compensate for the ICT demands placed on them. It is possible that the animation was somehow distracting to the high ICT group which meant their responses were not as well thought out.

Discussion and implications

The outcomes indicate that there was little effect on quantitative measures of item difficulty when the item format was changed. Even when the effect of ICT competence on item difficulty was examined, there was very little difference amongst the subgroups. The exception was the anomaly relating to the 'ICT High' group, where the item with animated stimulus appeared to be more difficult than the item with a static stimulus.

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. The sample consisted of 15-year-old students, who may have a very high level of digital literacy compared to the population and this needs to be considered when evaluating the outcomes of this study. There is evidence that poor item design has an impact on the validity of test scores (Huff and Sireci, 2001); therefore, it is important to establish if any of changes to item format would constitute 'poor item design' which would make a 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. In particular, item design may not have a noticeable effect on the average score in a class of students, but it is possible that individual students may respond very differently to a specific item design.

Limitations

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. Further investigation could be carried out to look in more detail at the possible confounding effects of ICT competence in this area.

The implications of the study are that the measures of item difficulty appear to be relatively unaffected by the item format presented to the student. However, the computer-based assessments in this study were undertaken by a sample that may have a high level of competence in computer-based applications relative to the general population and this may have affected the findings.

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, 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, 593-602.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, **32**, 2, 159-181.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, **48**, 2, 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, 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, 16-25.

Kane, M. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. 4th ed. 17-56. Westport, CT: Praeger.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, **20**, 8, 15-21.

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

Messick, S. (1989). Validity. In: R. L. Linn (Ed.), *Educational Measurement*. 3rd ed. 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, 475-479.

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

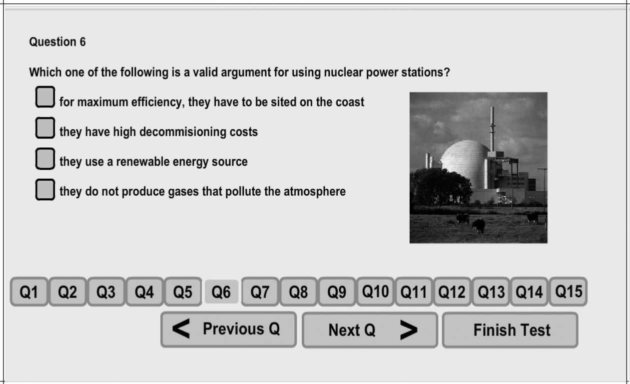
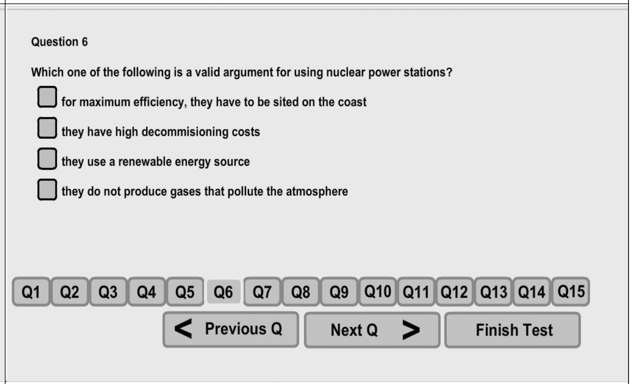
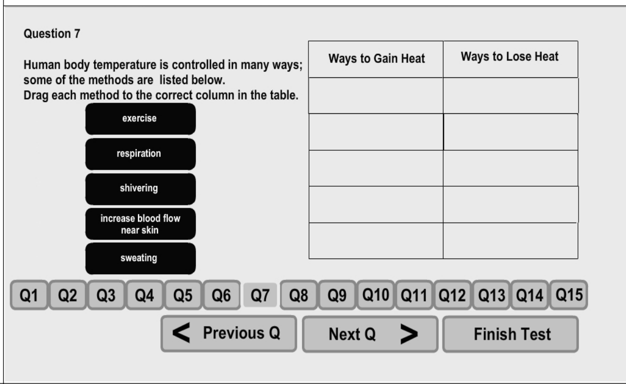
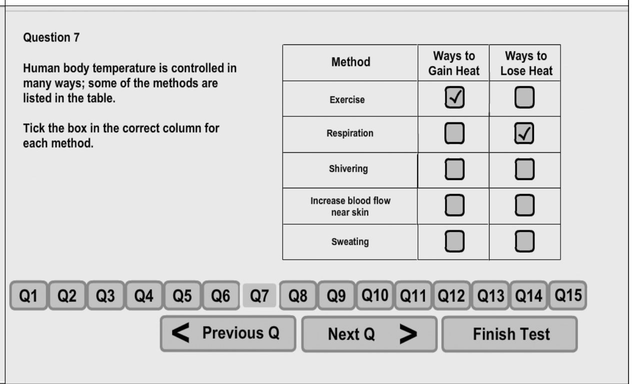
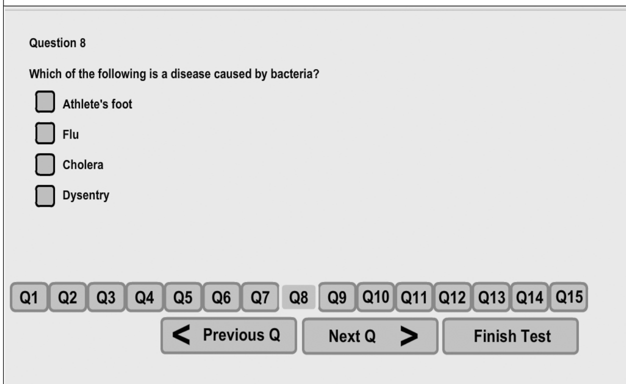
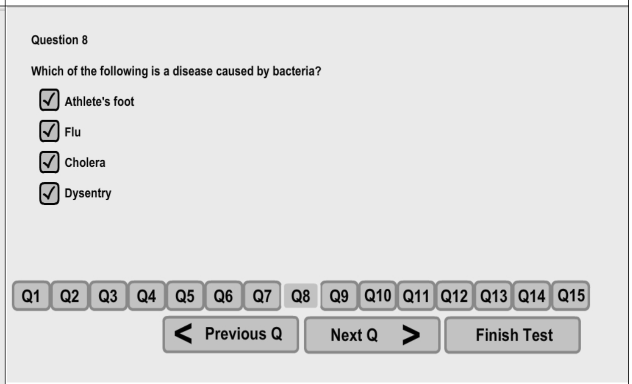
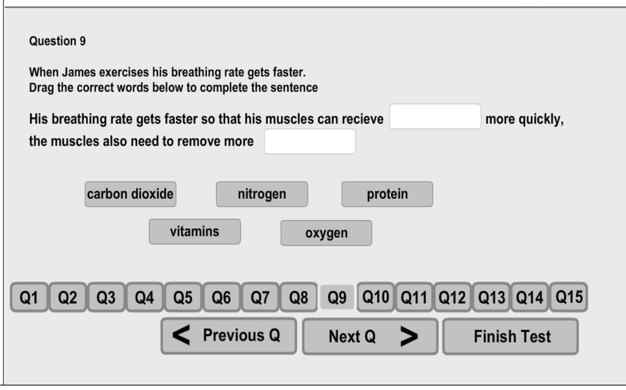
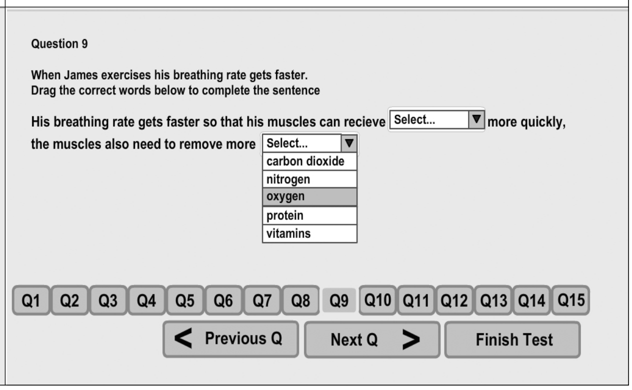
Russell, M., Goldberg, A., & O Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, **10**, 3, 279-293.

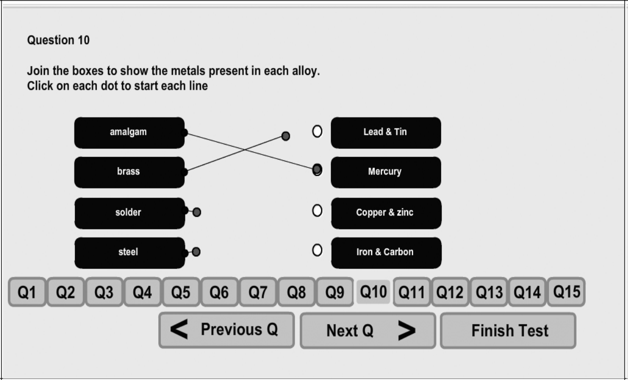
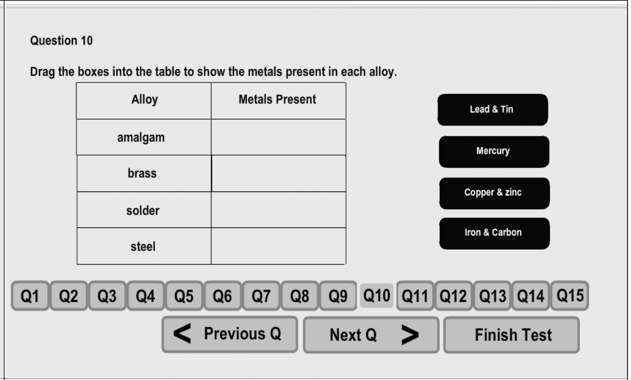
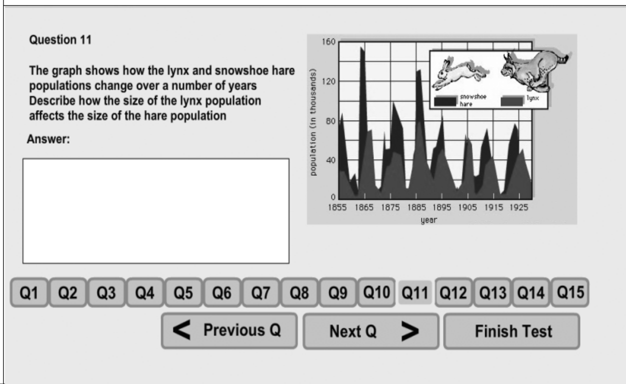
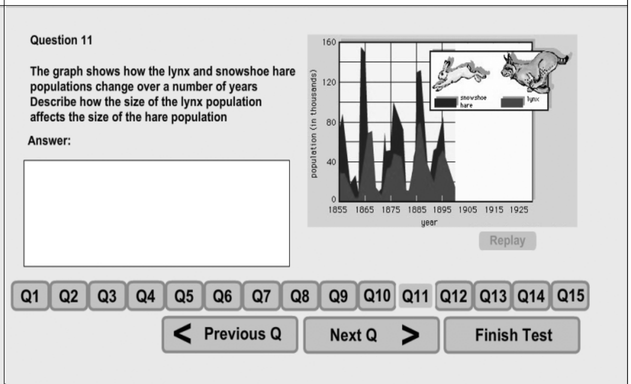
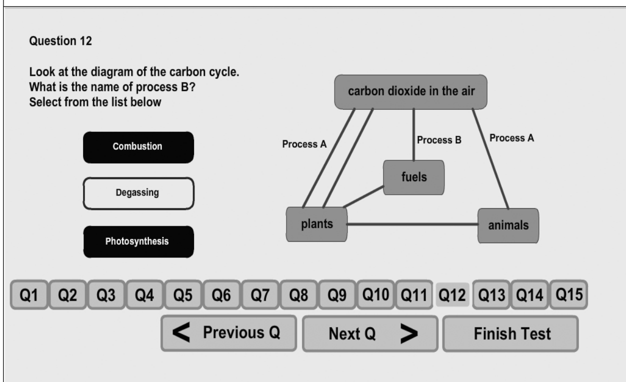
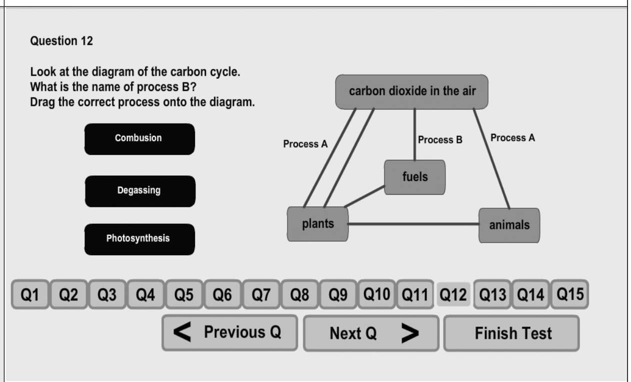
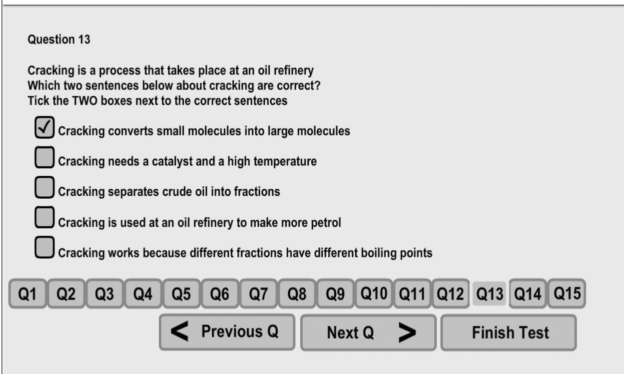
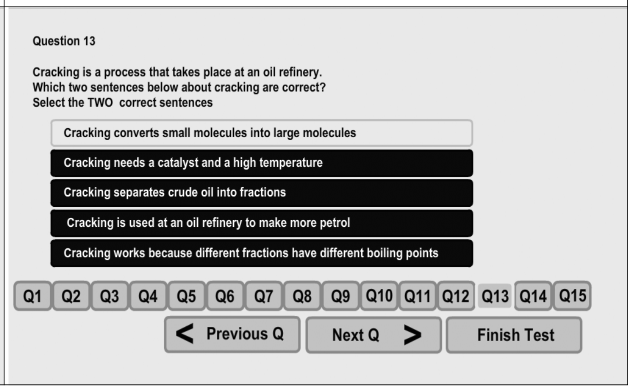
Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In: S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. 329-347. Mahwah, NJ: Lawrence Erlbaum Associates.

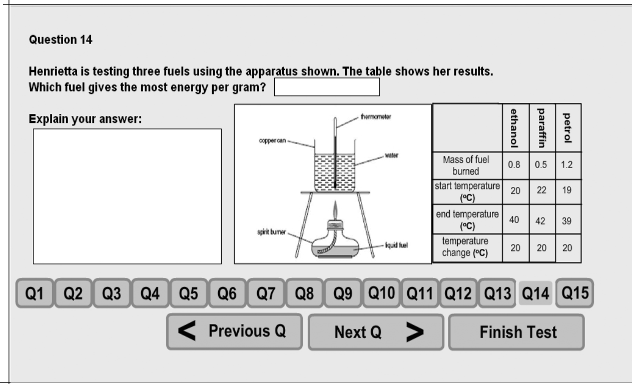
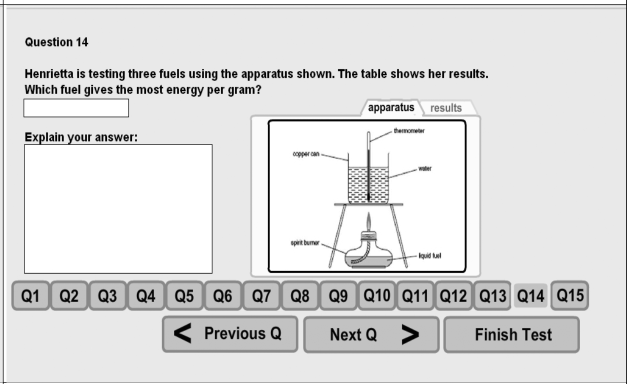
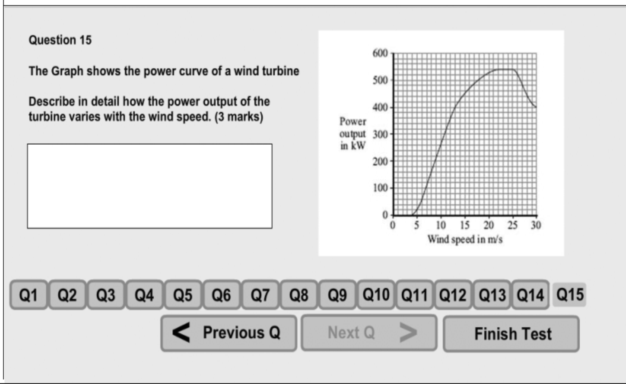
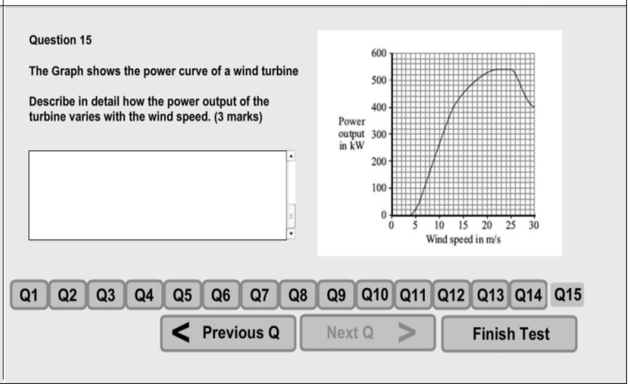
Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide*. Cambridge, MA: MIT Press.

Appendix: Modified test items

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

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

<table>
<tr><th></th><th>Form 1</th><th>Form 2</th></tr>
<tr><td>Q10</td><td>Matching options with lines</td><td>Matching objects, drag and drop into table</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
<tr><td>Q11</td><td>Static graphic</td><td>Animated graphic with replay option</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
<tr><td>Q12</td><td>Select option response</td><td>Drag answer to target response</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
<tr><td>Q13</td><td>Tick-box multiple choice</td><td>Selected statement multiple choice</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
</table>

<table>
<tr><th></th><th>Form 1</th><th>Form 2</th></tr>
<tr><td>Q14</td><td>All required data presented on one screen</td><td>Required data accessed by tabbed panels</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
<tr><td>Q15</td><td>Limited text box input</td><td>Unlimited text box input with scroll-bar</td></tr>
<tr><td></td><td>
</td><td>
</td></tr>
</table>