Centre for Education Research and Policy

THE COMPARABILITY OF ONSCREEN AND PAPER AND PENCIL TESTS: NO FURTHER RESEARCH REQUIRED?

Christopher Wheadon

SUMMARY

This paper presents an analysis of the first national high-stakes onscreen assessments offered in the United Kingdom, and considers the research and policy agendas required to support their development and use alongside paper and pencil alternatives. It argues that the UK has a lot to learn from the United States in the area of comparability, as the onscreen assessments currently being launched in the UK are far closer to US high-stakes assessments than the paper and pencil tests ever were. The US now has the benefit of twenty years of comparability research on such tests and has conducted an extensive search for so called test mode effects which affect the comparability of the results from two identical tests with different administration methods. Two findings in particular are unequivocal: non-speeded objective tests are rarely liable to test mode effects, and when they are liable to these effects, they are marginal; test anxiety caused by poor software design or unfamiliarity with the assessment environment can have a deleterious effect on results. It does not seem sensible or necessary to replicate these findings in the UK context; rather resources should be utilised to move onscreen assessments beyond their current conceptualisation. As the present study found a marginal test mode effect due to speededness, the paper considers a regulatory framework that should be put in place to monitor the introduction of the first generation of high-stakes onscreen assessments in the UK. In brief, the assessments should show evidence that onscreen tests do not offer undue advantages over their paper and pencil equivalents due to the speed with which they can be answered, and that suitable practice tests should be made available to reduce test anxiety. A light regulatory framework, rather than the onerous requirement to demonstrate comparability of every test introduced, should stimulate the innovation required to develop onscreen assessments that encourage and reward construct relevant behaviour.

INTRODUCTION

High-stakes onscreen testing in the United Kingdom is still very much in its infancy, involving automation of assessment tasks rather than their reconceptualisation (Bennett, 1997). The emerging high-stakes onscreen tests in the UK are online translations of their paper and pencil equivalent, and the regulator of these tests, the Qualification and Curriculum Authority (QCA) is preoccupied with the equivalency between the online translation and its paper and pencil equivalent. While the fact that assessment in the UK appears to be lagging ten years behind the United States may wound national pride (ten years may be an underestimation as Bennett’s description of first generation of tests refers insouciantly to item banks, on-demand testing and instant results services, and the first meta-analysis of onscreen tests appeared in 1988) there should be a competitive advantage to be gained. There should be no need, for example, to repeat the research that has underpinned the first generation of high-stakes tests in the US; this should reduce development and implementation costs. While the assessment

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales (company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX

systems of the two countries are fundamentally different in a number of respects, the first generation of onscreen assessments, constrained by limited item formats, brings them closer together than ever before. It would seem wise therefore to carefully consider the research findings on equivalency from the US before proceeding too far with the introduction of a regulatory framework and programme of research for onscreen assessments in the UK.

Fast forwarding through the research to the regulatory framework currently in operation in the United States does not, however, offer much comfort:

> "If a test is designed so that more than one method can be used for administration or recording responses—such as marking responses in a test booklet, on a separate answer sheet, or on a computer keyboard— then the manual should clearly document the extent to which scores arising from these methods are interchangeable" (American Educational Research Association, 1999, p. 70).

It would seem that the American Educational Research Association (AERA) considers that it cannot simply be assumed that the test scores from two administrations using different modes are equivalent. Bugbee (1996) reached this conclusion in his review of the comparability literature, stating that onscreen and paper and pencil tests can be equivalent, but it is the responsibility of the test developer to show that they are. This view has been picked up in the UK, in that it is considered necessary to establish score and construct equivalence if scores of paper and pencil and onscreen tests are to be interchangeable (MacDonald, 2001). Such a requirement, that equivalence must be actively demonstrated, would sound the death knell for high-stakes onscreen assessment in the UK.

High-stakes assessment in the UK is undertaken by a number of private organisations of charitable or profit-making status known as Awarding Bodies. Between them they assess a bewildering array of syllabuses ranging from Dance to Physics at a number of different levels, often more than once per year. The assessments are thorough examinations of the syllabuses that have been taught and use assessment methods designed to suit the particular construct under examination. The cost of proving the equivalence of scores for onscreen and paper and pencil versions of such a varied assessment diet could not be borne by the awarding bodies, and would likely result in an end to innovation. UK high-stakes onscreen assessment would never move beyond the first generation.

It seems that the need to actively demonstrate the equivalence of scores is a result of an apparent lack of consensus in the literature on whether test mode effects exist. The first major review of the literature yielded a number of detailed findings, only to conclude there were no generalisations to be drawn (Mazzeo & Harvey, 1988). The next major review found three studies that showed onscreen tests yielded higher mean scores than paper and pencil tests, nine studies in which computer based tests had lower mean scores than paper and pencil tests, and eleven studies in which no difference was found (Bunderson, Inouye, & Olsen, 1988). Once again, a number of notable characteristics of the incidence of test mode effects were noted, but the conclusion drawn seems to have been that you have a fifty per cent chance of achieving equivalent results between paper and pencil tests and their onscreen equivalents (Bugbee Jr., 1996; Clariana & Wallace, 2002). This extrapolation, however, ignores the characteristics of the tests and the test takers involved. Wise and Plake's (1989) subsequent review identified particular causes for concern, and once again called for separate norms for different modes of test delivery.

Since these early reviews the characteristics of the tests and the test takers have been analysed in great detail in search of what are now termed test mode effects: variables that could influence the comparability of onscreen and paper and pencil tests. In terms of the characteristics of the test takers, gender, ethnicity, socio-economic status, test anxiety, familiarity with computers, ability, attitude, effort and endurance have all been scrutinised. In terms of the characteristics of the tests, the test content, item format, flexibility of the software, size and colour of font and the speededness of the test have received an equal amount of attention. Given such a large number of variables, it is highly unlikely that some would not interact to result in test mode effects – leading to the natural conclusion that it is unsafe to generalise about test mode effects from one test to another, and the requirement therefore that comparability be proved or separate norms be published.

There are, however, certain findings that seem to be uncontroversial and that can lead to the production of general guidelines for the UK, without the need for unnecessary and costly research to establish comparability at a prohibitive cost. These findings are:

1. Speeded tests are liable to substantial test mode effects (Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; Pomplun, Frey, & Becker, 2002; Wise & Plake, 1989).
2. Non-speeded short answer or objective tests, regardless of item format, are not liable to substantial test mode effects (MacCann, 2006; Mead & Drasgow, 1993; Wang, Jiao, Young, Brooks, & Olson, 2007).
3. Test anxiety, caused by unfamiliarity with the testing environment, inflexible software or computer anxiety can adversely affect test results (Kveton, Jelinek, Voboril, & Klimusova, 2007; Smith & Caputi, 2007).

Great efforts have been made in the UK to remove time pressure from assessments as it is considered a construct irrelevant difficulty. Furthermore, the onscreen assessments currently available are confined to objective test formats. As for test anxiety, this may appear to require research specific to each assessment area, but as long as re-take opportunities on paper and pencil equivalents are available, candidates who feel they have underperformed in one format can take an alternative format. This does not, however, excuse the test developer from causing unnecessary test anxiety through poorly developed software. It would seem to suggest that practice tests should also be made available so that candidates can make an informed decision about the test mode they would prefer to undertake before entering a high-stakes testing situation. Following the launch of high-stakes onscreen assessment in the UK in 2006, therefore, it would seem propitious to analyse whether or not the tests that were delivered during the launch were speeded, as this would seem to pose the greatest risk of incurring a test mode effect.

Consequently, the question to be answered is whether the onscreen high-stakes tests offered in the UK are speeded and, therefore, liable to a test mode effect. This would be a simple question to answer in laboratory conditions, but any artificial situation could not claim to reproduce the speed of answer that candidates produce in an examination that they have been carefully prepared for and which has serious consequences for their educational career. In contrast, the operational data has the benefit of being produced in live conditions, but it is limited as it offers no item level timing data and, for the onscreen tests, no distinction between wrong answers and questions that have not been attempted.

METHOD

The analysis was undertaken on the data from the first onscreen high-stakes national examinations to be offered in the UK, the AQA GCSE Science syllabus. These examinations are designed so that they are available at two levels: higher tier and foundation tier. For candidates entered for the foundation tier, grades C to G are available. For candidates entered for the higher tier, grades A*−D are available. The higher tier tests consisted of nine sections with each section beginning with a stimulus of some kind in the form of a paragraph of text or a graph or series of diagrams accompanied by a description of some kind. Two of the sections required candidates to match four pieces of information, successful completion of which was worth four marks. The remaining seven sections consisted of four multiple choice questions, worth one mark each, which were related directly or indirectly to the stimulus. The foundation tier consisted of six matching questions worth four marks each followed by three multiple choice sections, again worth four marks each. An example of each type of question is included in Appendices 1 and 2.

The analysis was conceptualised as follows. Firstly a consideration of predicted outcomes and actual outcomes by mode can be analysed to highlight whether it was likely that a test mode effect occurred. Should candidates who undertook the tests onscreen perform better than would be expected from their prior achievement profile then this may suggest they have been advantaged. Secondly, item facilities could be compared and a DIF study undertaken. The model used for DIF was the One Parameter Logistic Model (OPLM) (Verhelst, Glas, & Verstralen, 1995) which, by allowing item discrimination values to differ, is essentially a two parameter model. This model therefore allows detection of DIF at different ability levels (non-uniform DIF). The methodology for the DIF detection was as follows. A random sample of 10,000 candidates was taken for each subject from the paper-based group, this being the maximum number allowed due to software restrictions on the number of cases that can be analysed. Any questions with negative item *x* total test score correlations were excluded as were questions that grossly misfit the OPLM. Item discrimination values (α) were then derived for the remaining items and item information functions analysed to compare the relative performance of matched ability groups in the two groups on every item. A test-mode effect caused by the speeded nature of a test should reveal questions becoming differentially easier for the onscreen mode towards the end of the test.

RESULTS

Table 1 illustrates the number of candidates who undertook the first round of examinations in each of the modes. The low numbers dictated that an item level analysis of Biology and Physics at foundation tier would appear sensible.

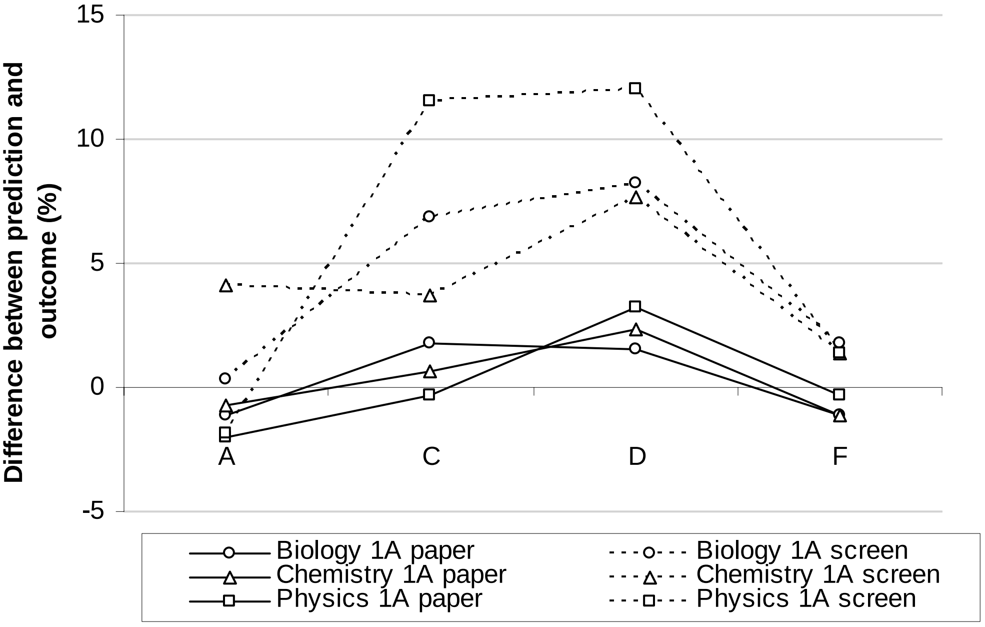
Table 1: Candidate numbers for GCSE Science A November 2006

Mode	Biology 1A (B1A)		Physics 1A (P1A)		Chemistry 1A (C1A)	
	Foundation	Higher	Foundation	Higher	Foundation	Higher
Paper	46803	57873	31787	42540	36410	48192
Screen	304	84	160	90	117	43

Is there a test-mode effect?

Figure 1 shows that candidates with the same level of prior achievement (mean Key Stage 3 score) achieved higher grades on the onscreen test than on the paper-based test in all three subjects. This may suggest a test-mode effect, although the low numbers involved mean that these figures require cautious interpretation. It could be the case that schools who are early adopters of the onscreen tests may differ structurally, in the quality of teaching or hours devoted to this particular syllabus for example, and the results are in line with their expectations.

Figure 1: The relationship between prior attainment (mean Key Stage 3 score) and outcomes for Biology 1A, Chemistry 1A and Physics 1A.



Item facilities

Should the items be performing consistently between the two test modes, the plot of item facilities should be parallel. Where they deviate from this, DIF may be suspected. The predicted outcomes for candidates based on their prior achievement (Figure 2) would suggest that the item facilities for the paper-based mode would be higher than the onscreen mode, although only marginally so for the Physics examination.

Figure 2: Predicted outcomes for foundation tier Biology and Physics based on prior achievement (mean Key Stage 3 scores)

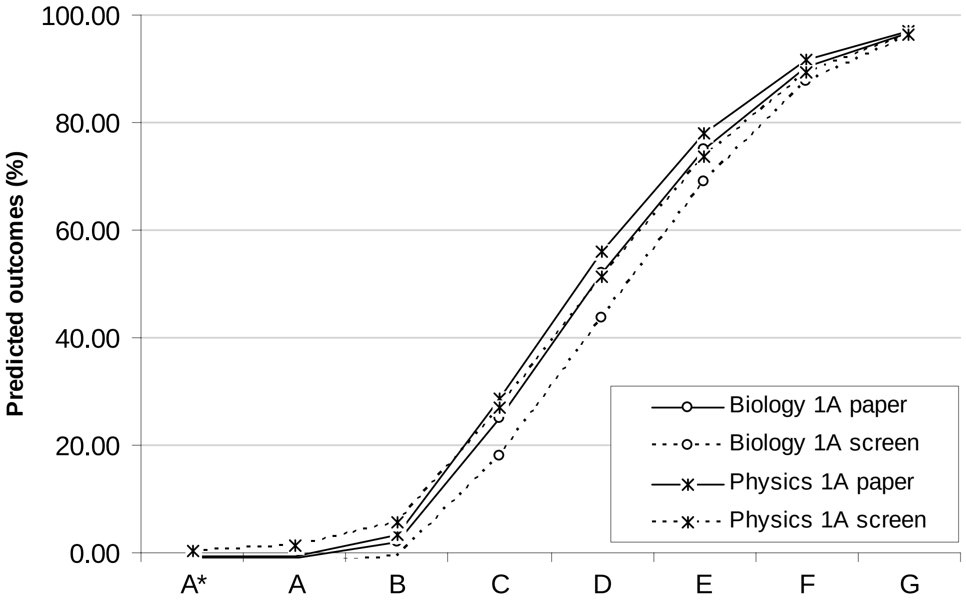
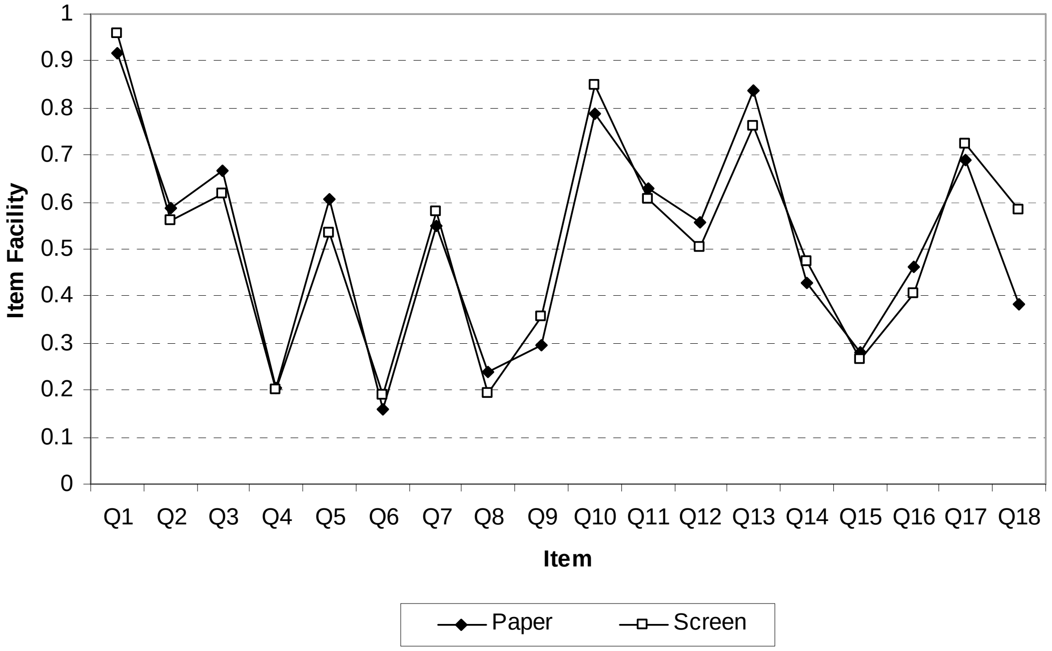


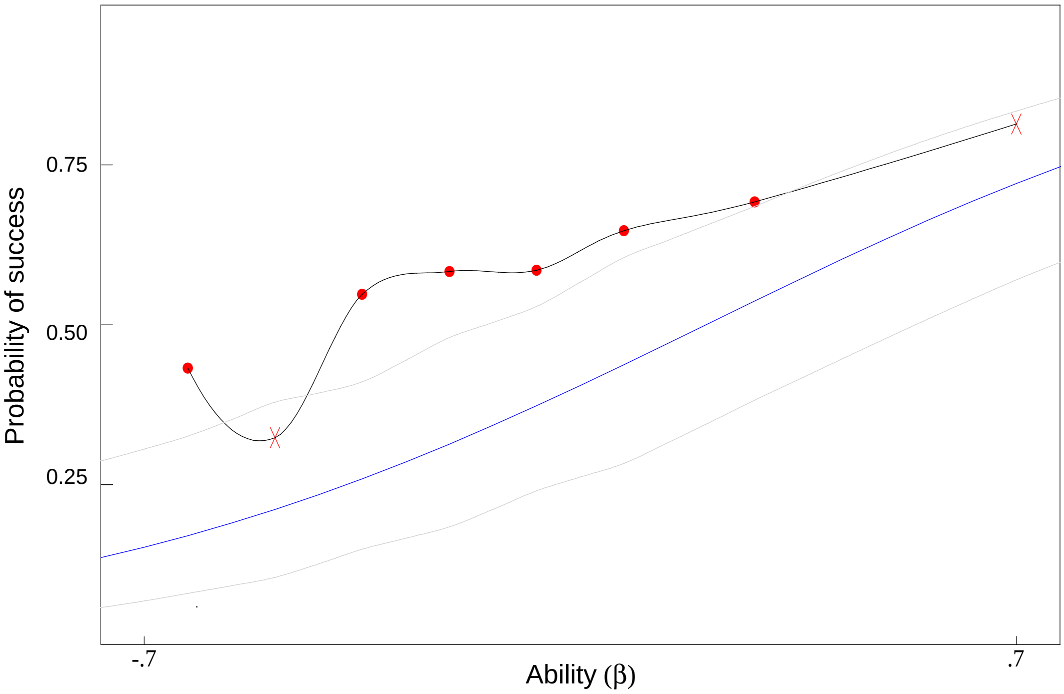
Figure 3 shows that, for the Biology foundation tier, contrary to expectation, some items had a higher item facility for the onscreen group than they did for the paper-based group. On these questions the onscreen group outperforms the paper group, but there is little suggestion of a systematic timing effect as the questions they find relatively easier are located throughout the test.

Figure 3: Item facilities in test order for the Biology foundation tier



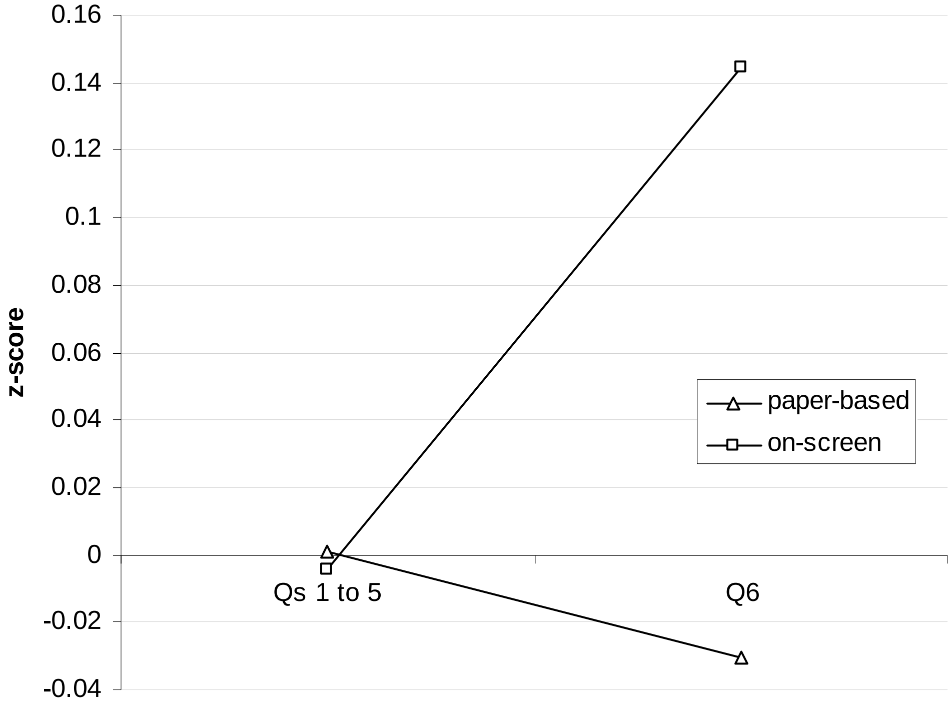
The last item on the Biology foundation tier, however, appeared substantially easier for the onscreen group. The DIF analysis confirmed that this item was indeed significantly differentially easier for six out of the eight score points modelled. Figure 4 illustrates the modelled item performance from the paper-based group (the darker central line), its surrounding confidence intervals, and the empirical performance of the onscreen group (marked by dots where it exceeds the confidence intervals). Candidates from the onscreen group had a consistently better chance of achieving correct responses on this question at all levels of ability. The effect would appear greatest for candidates of an average ability. This would be intuitive: the brightest would complete the test within the time limits on either mode, while the weakest would not be able to capitalise on any timing advantage.

Figure 4: Item Information Function for Question 18 B1A foundation tier.



Discounting the possibility that there was something peculiar about the design of this item that made it particularly easy onscreen, it was conjectured that this could be a timing issue. If candidates were performing faster on the onscreen test they would have a higher chance of success on the final questions. Figure 3 shows the onscreen candidates outperforming the paper-based candidates on a number of the later questions, with the notable exception of question 16. To investigate this issue further, the scores of candidates were split into their score on the first five sub-sections and their score on the last sub-section. These scores were then standardised to allow direct comparison. A two-way, mixed ANOVA was then conducted on each level of mode, the prediction being that there would be a significant interaction between mode and question positioning. The graph of means (Figure 5) suggested that this was indeed the case: although the onscreen group performed less well generally than the paper-based group on the first five sections; on the last section they outperformed them. While the interaction was statistically significant, however, the effect size was small (F(1,10305) = 8.453, p<.05, η^2 = .001). Indeed, after converting the standardised data of Figure 4 back into marks, the improvement on the last section is no more than a fifth of a mark. It should be remembered, however, that the non-uniform DIF suggests that the advantage is only reaped by candidates of middling ability: on the last question such candidates had a 30% higher chance of success than would be expected from their overall ability.

Figure 5: Mean score on the first 5 sections compared to the last section by mode, B1A foundation tier.



Turning to the Physics foundation tier it was then considered whether a similar interaction would be apparent. Unfortunately the last question on the Physics examination showed a negative test-item correlation which meant that it could not be modelled. The two questions directly preceding it, however, both showed evidence of non-uniform DIF: these questions were differentially easier for lower ability candidates. Once again there was the necessarily significant interaction between mode and sub-section (F(1,10160) = 5.599, p<.05, η^2 = .001) but once again the effect size was small.

DISCUSSION

There is no doubt that the results from the first set of onscreen examinations in the UK set off some alarm bells, as candidates for the onscreen version of the test outperformed their predictions. The schools taking the onscreen test, however, were self-selected early adopters who may differ systematically from the majority group. This study shows a test mode effect due to timing of one fifth of a mark. This finding cannot obviously be generalised to all assessments that are planned for onscreen delivery, so the following warning from AERA should be heeded:

> “A shift from paper and pencil format to a computer-administered format may affect test speededness” (1999, p. 33).

As a consequence, procedures should be put in place to ensure that candidates are not advantaged by a change in speededness. What of a test mode effect caused by software design, item format, or some other variable? The evidence from the US is clear: the test mode effect is small for non-speeded objective tests.

While the scope of this analysis was narrow, and it is hard to generalise from the conclusions, it would seem perverse not to learn from the lessons in the US. The following points, therefore, are a draft of the regulations that should be put in place to ensure comparability of the first generation linear onscreen assessments:

A. If assessments consist of objective test or short answer formats:

1. On first examination, evidence should be provided on item completion rates from onscreen and paper assessments to show that they require candidates to work at a similar pace.
2. Practice tests should be available to all candidates to alleviate test anxiety caused by unfamiliarity with the testing environment.

B. Where assessments depart from objective test or short answer formats evidence will be required to show that construct irrelevant variance is not introduced by the method of assessment.

While it may seem that item B may act in favour of the onscreen mode of assessment, it would seem be counter-productive to restrict innovations in assessment that allow candidates to demonstrate more fully their ability in a particular subject.

While onscreen tests remain in the first generation, research effort should not be expended on replicating the findings of comparability research that have been undertaken in the US, but should focus on moving UK assessment into its second generation; reconceptualising assessment in a way that encourages construct relevant behaviour on the assessment task. UK high-stakes onscreen assessment may currently be 20 years behind the US; to insist on a programme of comparability research would condemn our candidates indefinitely to a mode of assessment that may soon seem anachronistic. The guidelines on comparability outlined here would ensure that further research is not undertaken needlessly.

Christopher Wheadon
4 December 2007

Appendix 1: A multiple choice section from the Biology examination

QUESTION SIX

Digitalis is a toxin which is extracted from plants such as foxgloves.

Digitalis can be used to treat patients who are likely to suffer from heart failure. Digitalis affects the rate of heartbeat and the volume of blood pumped per heartbeat.

The table shows the effect of using different concentrations of digitalis on the heart action of a male patient.

Concentration of digitalis in arbitrary units	**Mean rate of heartbeat in beats per minute**	**Mean volume of blood pumped per heartbeat in cm^3**
0	136	35
10	120	46
20	103	54
30	71	59
40	59	62
50	47	63

6A If 20 arbitrary units of digitalis were used on this patient, the amount of blood pumped by his heart, at rest, would be…

1 1.91 cm^3 per minute
2 5.15 cm^3 per minute
3 2060.0 cm^3 per minute
4 5562.0 cm^3 per minute

6B Which one of the following best describes the effect that increasing the dose of digitalis has on the activity of the heart?

	Effect on heart rate	**Effect on volume of blood pumped per beat**
1	increase	increase
2	increase	decrease
3	decrease	decrease
4	decrease	Increase

6C It would be unsafe to use the results from this patient to decide the dose for other patients because…

1 digitalis has not been trialled on human volunteers
2 side-effects may harm the patient
3 the sample size is not large enough to draw conclusions
4 drug companies may put undue weight on the results from the first patient

6D Which of the following best describes the term ‘toxin’?

1 a chemical that affects the heart
2 a poisonous substance
3 a substance produced by a plant
4 a useful drug

Appendix 2: A matching question from the Biology examination

This question is about the possible effects of some substances on the body.

Match effects, A, B C and D, with the numbers 1-4 in the table.

A may cause damage to the liver and brain
B may lead to heart disease
C may lead to becoming addicted to hard drugs
D may lead to a reduced level of cholesterol in the blood

	Effect may be caused by …
1	Drinking large amounts of alcohol
2	Eating food containing a lot of unsaturated fats
3	Eating large amounts of low density lipoproteins
4	Taking cannabis

REFERENCES

American Educational Research Association. (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Bennett, R. E. (1997). *Re-inventing Assessment*. Educational Testing Services.

Bugbee Jr., A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*, 282–299.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

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

Kveton, P., Jelinek, M., Voboril, D., & Klimusova, H. (2007). Computer-based tests: the impact of test design and problem of equivalency. *Computers in Human Behavior, 23*(1), 32-51.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79-91.

MacDonald, A. S. (2001). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39*, 299-312.

Mazzeo, J., & Harvey, A. L. (1988). *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests: A Review of the Literature* College Board Publications: College Board Report No. 88-8.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive-Ability Tests - a Metaanalysis. *Psychological Bulletin, 114*(3), 449-458.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Smith, B., & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior, 23*(3), 1481-1498.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One-parameter logistic model: OPLM. Arnhem: CITO.

Wang, S. D., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219-238.

Wise, S. L., & Plake, B. S. (1989). Research on the Effects of Administering Tests via Computers (Vol. 8, pp. 5-10).