

**Consensus statement**  
**on technical issues**  
**in the early stages of using e-assessment**  
**in UK general qualifications**

## **Introduction**

1. This is a consensus statement issued on behalf of the e-assessment Technical Working Group (e-aTWG). e-aTWG is composed of researchers and e-assessment managers from UK regulators and GCE and GCSE Awarding Bodies. Group members' names are at the end of this document.

### ***The nature of consensus statements***

2. Consensus statements are widely used in medical research and practice. Consensus development also has an important role in regulatory practice.
3. Consensus statements address issues upon which research evidence is not yet complete. In such statements, signatories synthesise available evidence and state their joint view (the consensus). They also recommend next steps to take. The point of doing this is both to make current understandings plain and to ensure that future practice is grounded in sound principles.

### ***Evidence base for this consensus statement***

4. This statement summarises the work of the e-aTWG in its meetings between November 2007 and January 2009. In doing so, it is based on: reviews of e-assessment literature, outputs of scenario planning exercises and reports of empirical research carried out following the January and June 2008 examinations series.

### ***Definition of e-assessment***

5. For the purposes of this statement, e-assessment means an assessment in which the student interacts with the exam questions or other assessment materials via a computer. That computer could take several forms, such as a desktop, laptop or handheld computer. Other uses of technology to support examinations processing – such as onscreen marking, remote (technology-supported) standardisation of markers, or remote awarding meetings – are excluded from this definition of e-assessment.

### ***Extent of use of e-assessment***

6. e-assessment is not really a new thing. It has been around for at least 20 years and in some contexts – e.g. the USA, some professional qualifications or licensure examinations in the UK – it is widely used.

7. However, e-assessment is still quite sparsely used in assessment for general qualifications. Only a small number of GCSEs and GCEs are accredited to use e-assessment. Also, typically only a small element of a particular qualification will be accredited to use e-assessment. Further, at the present time, there will generally be small cohorts of candidates sitting each e-assessment.
8. The information above suggests that e-assessment is still in the early stages of uptake in GCSEs and GCEs.

## **Comparability**

9. Some e-assessed qualifications are entirely new, with no existing pencil-and-paper (p&p) forerunner. Such entirely new qualifications would typically have relatively small cohorts. The other scenario is that e-assessment is introduced into an existing qualification, often as an option, and sits in parallel to an existing p&p assessment within a qualification. This latter scenario affects both qualifications with large cohorts and those with high stakes.
10. The following table shows some example qualifications that are accredited to use e-assessment from the five GCSE and GCE awarding bodies:

<b>Awarding body</b>	<b>Qualification type</b>	<b>Qualification title</b>
AQA	GCSE	Science A
CCEA	GCE	Moving Image Arts
Edexcel	GCSE	Construction and Built Environment (Pilot)
OCR	GCSE	Environmental and Land-Based Science
WJEC	GCE	Applied ICT

11. When two modes of assessment (p&p and e-assessment) are available as parallel options within a qualification, the matter of comparability crops up. A definition of comparability for this context will be a substantive element of the consensus to be output from this paper. Users of qualifications such as Higher Education, employers and government are entitled to be indifferent as to the mode by which a candidate was assessed. They are entitled to treat grades from different modes as being interchangeable.

## **Definition**

12. Comparability is a major issue in research into examinations in the UK, and there are many definitions of comparability, and conditions under which such definitions operate.
13. We need a practical definition of comparability that will be suitable for the context in which UK awarding bodies and regulators work, and the conditions described

below. In this context, test outcomes can be said to be comparable if features (such as central tendency and spread) of the score and grade distributions of candidates on both modes appear to be similar for equivalent groups of candidates, taking into account other known evidence about the test takers' knowledge, skills and understanding (KSU) (such as information about their prior attainment).

14. The present definition operates when tests are strictly parallel in content, item type and so on; that is, the only substantive difference between two tests is the mode of delivery (p&p versus on-screen). In order for tests to be parallel within the current definition, one should be able to establish that the tests in the two modes assess the same KSU. Further, when talking about comparability in this context, one is talking about candidates' KSU assessed by the test in question, at the time that the test was taken; not their potential to do well in the future or their general intelligence.
15. This definition of comparability is an essentially statistical one. It may be reiterated that it applies for the purposes of this consensus statement. Other aspects of the examinations process (such as the setting of grade boundaries and the maintenance of exams standards) require the application of professional judgement in the light of all the evidence.

### **Existing research evidence about comparability**

16. Comparability has been one of the most frequently tackled issues in e-assessment research to date. Many studies (and also consolidating meta-analyses) have shown that, when the operating condition of strictly parallel tests (as defined above) applies, comparable scoring (as defined above) can be achieved between e-assessed and p&p modes. These findings typically apply to 'simple' item types, such as multiple-choice or short-answer questions.
17. Further, in many cases where there have been found to be significant differences in scoring between e-assessed and p&p modes, the effect size has been small. To put it another way, there is a real difference in scoring, but it is only a small difference in effect.
18. There can be lack of comparability when certain conditions apply. Important factors associated with lack of comparability between e-assessments and p&p tests would include: differential speededness (i.e. candidates are less pushed for time in one mode than in the other), candidates having problems with an e-assessment user interface (UI) and/or candidates feeling unfamiliar with an e-assessment interface and hence anxious. Researchers have suggested that the

last two threats to comparability can be overcome if candidates have sufficient opportunity to acquaint themselves with the UI via a practice test or similar utility.

19. There can be lack of comparability when comparability is defined more widely than in this paper. For example, some questions might have different difficulties when realised on paper or on screen.

### **Experimental conditions to investigate comparability**

20. In order to acquire useful findings, a researcher needs a well-defined and sufficient sample of candidates taking e-assessments and p&p tests to provide data from which to run analyses. Unfortunately, many e-assessments currently being run in GCSEs and GCEs only have small candidatures about which researchers do not know enough (e.g. the many factors that can affect exam performance, such as how well the candidates have been taught, how motivated they are, etc).
21. These problems with the samples upon which researchers can base findings should make those undertaking research cautious in interpreting test outcomes as comparable or not comparable.

### ***Test windows***

22. Test centres such as schools are likely to find it impractical to accommodate a large cohort of test takers sitting an e-assessment all at the same time. This is a problem for awarding bodies and regulators. There are various possible solutions to this practical problem. Such solutions include having a longer time period during which test sessions can be scheduled (a 'test window'), or having several (more than at present) single-day test windows throughout the school year.
23. There are advantages and disadvantages to having a few multiple-day test windows or relatively many single day test windows. There does not yet appear to be a consensus as to which is the better approach to alleviate logistical problems with running large cohort e-assessments.
24. The regulators have commissioned research to investigate best practice in 'on-demand testing'. This might provide more insight into issues here.
25. This issue is relevant at the current time, when most school computers are fixed desktop machines in IT rooms. The issue might become less relevant if and when schools have sufficient numbers of laptops, and reliable wireless networks are commonly available in schools.

## ***Prevention of cheating***

26. Qualifications need to be fairly gained. Cheating can happen in both p&p and e-assessment. e-assessment is under no greater or lesser obligation to prevent cheating than any other mode of assessment. Moreover, technology can be used to assist cheating in p&p testing as well as e-assessment.
27. Research into how to address technology-enabled cheating recommended integrating cheating prevention strategies by using the 'three e's' (ethics, engineering and enforcement).
28. There are specific cheating challenges that e-assessment faces that do not exist in p&p assessment. Firstly, it might be easier for candidates to see the text of exam questions and answers on screen (because screen text may be easier to see than text on paper, and/or because fixed desktop PCs are often closer together than desks in exam halls). Also, if exams are run across multi-day test windows, there is a risk that candidates will find out what is in the exam before the session is run.
29. Conversely, there are cheating-counteracting techniques (the engineering 'e' of the three 'e's) that are much easier to implement with e-assessment than with p&p tests. Such techniques include: providing test versions made up of different questions, changing the order of questions, changing the order of multiple-choice (MC) options within a question or parameterising values within MC options.

## ***Equalities issues***

30. General qualifications are covered by equalities legislation. Such legislation puts an active obligation on providers of public services to make sure that no candidates are put at a substantial disadvantage because of their disability.
31. An equalities issue might arise if an e-assessment provided a greater convenience for test takers. For example, if an e-assessment was available more frequently than its parallel p&p version (for example via a multi-day test window or multiple single-day windows) and certain groups of candidates could not access the e-assessment on grounds of faith (objection to technology) or disability, then such candidates might not have the opportunity to enjoy the benefits of e-assessment available to other candidates.

## Recommendations

32. When considering comparability between e-assessments and p&p tests, awarding bodies and regulators should understand the term to be defined by paragraphs 12 to 15, above.
33. It would be wise to pay heed to prior research findings when designing e-assessments to be strictly parallel to p&p versions. So, for example, it would be wise to avoid test versions where the modes of assessment were differentially speeded. (The best way to ensure this is to make ample time available, whatever the test mode.) Also, research evidence suggests that it is best to provide e-assessments in which the user interface does not present undue difficulties, and thus have a differential impact on candidates' test performance. Many other factors can also impact on comparability (see paragraph 18, above).
34. If e-assessments are well designed and minimise known sources of non-comparability, such as those referred to in paragraph 33 above, then e-assessment can be introduced to GCSEs and GCEs in advance of between-modes comparability being firmly established. Further, previous research evidence that suggests that e-assessment and p&p tests can provide comparable measurement is an important factor to be taken into account when deciding whether a new e-assessment is parallel to its p&p alternative.
35. If an e-assessment is introduced in advance of empirical confirmation that it is providing comparable measurement to its p&p alternative, then there must be a corresponding commitment to monitor this between-modes comparability and to timely information sharing about the outcomes of such monitoring. These mutual obligations could be expressed as the regulators' commitment to facilitating innovation and awarding bodies' commitment to the ongoing robust monitoring of comparability, and transparency and dialogue between awarding bodies and regulators.
36. Such monitoring of between-modes comparability is likely to be part of a programme of robust, critical and timely research into various aspects of e-assessment.
37. Whilst research based on weak samples will not be adequate to demonstrate comparability, it can be useful for trialling the techniques by which comparability can be proved as sample sizes and representation improve. Such initial work might also be a useful way to establish best practice in achieving samples of the necessary quality to conduct comparability research.

38. If sound research into between-modes comparability is designed and implemented from the start for e-assessed GCSEs and GCEs, it will be possible to make convincing statements as to the comparability of the e-assessed and p&p tests as soon as robust evidence is available. Further, this type of research will provide a sound basis from which to understand the measurement properties of more 'advanced' forms of e-assessment.
39. If p&p and e-assessed parallel tests turn out not to be parallel, then it seems best practice that comparability of grades should be treated as the more important objective, rather than comparability of scores. This is because grades are the reporting metric that matters most to candidates and test users. It might be possible for non-comparable scores to provide comparable grades.
40. The implications of multi-day test windows versus multiple single-day test windows need to be understood more thoroughly. Research into on-demand testing will go some way to resolving issues here.
41. There are physical ways to make cheating more difficult, such as placing guards on computer screens, barriers between workstations, etc. Where possible these should be used to inhibit cheating.
42. Engineering techniques to prevent cheating in e-assessments (changing the questions in tests, the order of questions, etc.) need to be further researched. Such research would come from two directions: do the techniques actually prevent cheating? And, what impact do such techniques have on comparability between versions?
43. Comparability remains a crucial factor in ensuring the public's confidence in GCSE and GCE examinations. As innovations in test designs, administration approaches, etc. flow from the increasing implementation of e-assessment, they need to be matched by a robust comparability framework.
44. Equalities implications of e-assessment need to be carefully considered. Impact assessments should be conducted to make sure that the use of technology does not increase inequalities in access to qualifications. However, it is believed that enhanced assessment opportunities could be provided by the introduction of e-assessment, and so the scenario described in paragraph 31, above, should not of itself prevent the introduction of e-assessment.

## Annotated bibliography

The comparability of p&p and e-assessment has been one of the most researched issues in e-assessment so far. There are many papers on the topic. The following examples are believed to be some of the most important and accessible. They are all available for free on the web at the time of writing.

Pamela Paek (2005) Recent trends in comparability studies.

[http://www.pearsonsolutions.com/downloads/research/TrendsCompStudies\\_rr0505.pdf](http://www.pearsonsolutions.com/downloads/research/TrendsCompStudies_rr0505.pdf).

Randy Bennett (2003) Online assessment and the comparability of score meaning.

<http://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf>.

Chris Wheadon (2007) The comparability of onscreen and paper and pencil tests: no further research required? <http://tinyurl.com/bnu4fh>.

The free online *Journal of Technology, Learning and Assessment* – [www.jtla.org](http://www.jtla.org) has lots of good articles about e-assessment, including several on comparability. The following are worth reading:

Martin Johnson & Sylvia Green (2006) On-line mathematics assessment: the impact of mode on performance and question answering strategies.

<http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1020&context=itla>.

Gautam Puhan and colleagues (2007) Examining differences in examinee performance in paper and pencil and computerized testing.

<http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1115&context=itla>.

Randy Bennett and colleagues (2008) Does it matter if I take my mathematics test on computer? a second empirical study of mode effects in NAEP.

<http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1140&context=itla>.

There are some regulatory documents relevant to this consensus statement. These include:

The regulatory principles for e-assessment. <http://www.ofqual.gov.uk/461.aspx>.

Jean Underwood (2006) Digital technologies and dishonesty in examinations and tests. <http://www.ofqual.gov.uk/177.aspx>.

The report 'principles and practice of on-demand testing' has been completed by AQA researchers and submitted to Ofqual in 2009. This should be posted on the web soon.

## Consensus statement signatories

Andrew Boyle

Principal research officer, and programme leader for e-assessment

Office of the Qualifications and Examinations Regulator (Ofqual)

[www.ofqual.gov.uk](http://www.ofqual.gov.uk)

Rose Clesham

Assessment design and e-assessment manager

Edexcel

[www.edexcel.org.uk](http://www.edexcel.org.uk)

David Crosbie

Principal officer – regulation

Council for the Curriculum Examinations & Assessment (CCEA)

[www.rewardinglearning.org.uk](http://www.rewardinglearning.org.uk)

Mike Forster

Head of Operational Research

OCR

[www.ocr.org.uk](http://www.ocr.org.uk)

Dale Hinch

e-assessment development project manager

Research & Design

Edexcel

[www.edexcel.com](http://www.edexcel.com)

Gwen Low

Chair of examiners

OCR

[www.ocr.org.uk](http://www.ocr.org.uk)

Alun McCarthy  
Head of ICT and Access, Qualifications and Learning Division  
Department for Children, Education, Lifelong Learning and Skills (DCELLS)  
<http://new.wales.gov.uk/topics/educationandskills/?lang=en>

Andrew Morse  
e-assessment development officer  
WJEC  
[www.wjec.co.uk](http://www.wjec.co.uk)

Dennis Opposs  
Head of Standards  
Office of the Qualifications and Examinations Regulator (Ofqual)  
[www.ofqual.gov.uk](http://www.ofqual.gov.uk)

Bob Penrose  
Programme manager  
Executive Management Division – Process Development  
AQA  
[www.aqa.org.uk](http://www.aqa.org.uk)

Raymond Tongue  
Head of Research and Development  
WJEC  
[www.wjec.co.uk](http://www.wjec.co.uk)

Chris Wheadon  
Principal research manager  
AQA  
[www.aqa.org.uk](http://www.aqa.org.uk)

Please address any queries or comments on this consensus statement to  
[info@ofqual.gov.uk](mailto:info@ofqual.gov.uk) who will ensure your query is directed appropriately.