International Conference
ICT FOR LANGUAGE LEARNING
7th Edition

Computerised Summative Testing: One step Forward or Two Steps Back?

Richard Chapman
University of Ferrara (Italy)
richard.chapman@unife.it

Abstract

The talk aims to initiate principled investigation into a subject of pressing concern. In recent years computerised systems of testing have become much more commonplace, with big testing institutions offering a computerised option, or even an exclusively computer-based test. The advantages of this in terms of delivery and, potentially, in terms of reliability are assumed and occasionally advertised unequivocally, while there has been little serious research into the real nature of this new test experience for the language learner and candidate. Sometimes even 'high-stakes' examinations are computer-based (at least as regards delivery) but the effects on candidate performance are not really understood. Is there an intrinsic difference to the computer-based English test, and is this difference linguistic, pragmatic or merely personal? Are the results liable to variation because of the instruments being used? What is the washback effect of a computer-based test as compared to traditional, paper-and-pencil English language examinations? The paper hopes to address these issues using the theoretical framework of the language testing literature and examples from current computer-based English language tests.

1. Introduction

Perhaps it is fair to say that CALT (computer-assisted language testing) has finally come of age. Recently, computer-based, or even on-line summative language tests have become widely available, with some of the principal players in the international certification market getting involved[1]. At last, high-stakes tests are being produced[1 and 2], in contrast to the plethora of placement tests, practice tests and the like which offered many of the benefits of computers without 'biting the bullet': using information technology to facilitate valid and reliable testing of language competence to certify achievement. If we really are convinced of the much-touted advantages of information technology in language learning, then the testing and certification community really ought to be making use of the tools available. This is all the more true if we accept that more and more language learners are exploiting on-line materials in order to acquire and practise new language: it may well be that in twenty years the current mismatch between computer-assisted learning (even with tablets and apps) and traditional paper-and-pencil tests still the norm today will seem unfair and almost absurd.

2. The advantages of CALT

Many advantages are claimed for computer-assisted testing which are worth summarising here [3]. Firstly, there are unquestionably benefits in logistic and administrative terms. Examinations need no longer be sent out by 'snail mail' and kept in a safe until the official time and date of the test, to be sent back as completed scripts to an official certification organisation for scoring, only for the results to be posted back to the candidate (or test centre) weeks or even months later. With CALT no paper need be shipped, and times can be drastically reduced. A second advantage is the claim of increased uniformity: the information for the test-taker during the examination will be identical for all (rather like the listening tests on cassette tape or cd in days of yore, when all the timing and administration of the test was effectively regimented by the recording which was never to be stopped). With CALT all candidate behaviour might also be accurately tracked (e.g. the time taken by each candidate on each section or item, the changes and corrections made etc.). Perhaps most importantly, there is the claim that CALT allows individualisation of the test according to level revealed by each response. This is known as CAT (computer-adaptive testing) where items are pre-identified as to difficulty and an algorithm dictates items the candidate must face depending on right-wrong decisions so far in the test. The benefit here is speed: an examination will have fewer redundant items (i.e. those far too easy or too difficult to tell the examiner anything significant about the candidate's real competence). CAT will also, it is claimed, provide a more accurate assessment of the test-taker's ability for this same reason of item selection, and it will probably result in a more pleasant test experience (because more suited to the candidate's level). There are also substantial advantages in terms of security: not only can we do away with the sealed envelopes not to be opened before the morning of the test, but more importantly, the actual tests will vary: each candidate will answer a different group of questions from the bank of thousands, ensuring that copying, or previewing of the papers, is impossible as each exam is unique (see OTE White Paper). Lastly, tests may be offered much more often over the year (e.g. the CELA Main Suite computer-based examinations), or even virtually on a demand basis in some cases (e.g. the OTE); gone are the days of two

sessions a year, and a failed examination meaning at least six months' wait before the next attempt. With results available quickly, if not instantly, and on-screen functions (such as a constantly displayed timer, and a *help* button), the benefits of CALT seem obvious.

3. CALT Caveats

Although the case for computer-assisted and even on-line language testing is highly attractive and has gained considerable ground over the past thirty years, there are issues serious enough to cause some concern. While it is unquestionably true that security can be improved by computer-assisted testing, the issue of candidate identity becomes foremost: identity detection of the actual test-taker using a computer is virtually impossible with present technology (this is described as "an insurmountable problem in high-stakes testing": Pathan, 2012: 40). The solution for this is usually to have recognised 'test centres' with proctors to verify identity and conduct during the test (as the OTE currently does). This can indeed reduce the gravity of the problem, but security does depend on the honesty of the recognised centre and the auditing carried out, and this comes along with a loss in terms of the very availability that is the great benefit of on-line testing. Clearly, you cannot do a high-stakes in your bedroom, at any hour of the day you might choose!

A second potential difficulty lies with the CAT function mentioned earlier and used in some, though by no means all summative CALT. Adaptive testing requires a well-constructed, pretested and evaluated item bank (see OTE white paper for a model of the procedure), but also one which is sufficiently large: if not, the security benefits of individualised tests are at risk. Candidates might easily 'get wise' to certain 'critical items' that define the level of later questions (and pass them on to other testees). This need not be an inevitable consequence of CAT, but it does remind us of the necessity of committing a great deal of time and resources into item bank construction, and insists on the need for regular renewal of items. It is clearly not a cheap option. Naturally, the algorithm must be nigh-on perfect, and the candidate must always be presented with enough items. Here we might also mention the risk of greater anxiety for the candidate in a high-stakes test where a certain item is not just worth the points awarded, but may also dictate the nature of other items to come.

There are significant technical issues as well. There is a certain 'constraint of medium' in all computer-assisted tests: the page is not as large as the classical A4, and so repeated scrolling up and down a reading passage is the inevitable result. It remains to be seen whether in the long-term the effect is a reduction in the length of reception instruments[2] or a change in their characteristics. At the same time, it is necessary to remember that CALT requires standardisation of equipment and adequate technical expertise to be on hand during the administration of the test. Here there is always a threat of the 'digital divide' having an undue effect on candidate performance in what are very often international examinations (and widely sold as such). We should note also that even a reliable source of electricity is not a 'given' in many parts of the world.

4. One step forward, two steps back?

The caveats briefly outlined above can hardly be said to make the case against computer-assisted or on-line testing. They might act as warnings to those responsible for producing high-stakes, summative tests, but in a sense it has always been true that a placement test could be 'quick and dirty' and yet still operate reasonably effectively and without harming anyone, and a progress test, created largely for pedagogical purposes, might contain errors and yet not do serious damage, while summative testing required greater precision and professionalism, being as it was a shibboleth of some (possibly life-changing) importance [4]. It is no surprise that CALT has only recently 'come of age' and begun to play a part in summative language testing. While CELA started offering Main Suite examinations in computer-based form in 2006/7 (simply as an identical alternative to their paper-and-pencil tests: see *Research Notes* 22 [5]), 2012 saw the introduction of Oxford's OTE that is not simply computer-based, but fully on-line. A step, or perhaps a leap forward is being taken, with all the potential benefits in administration that we outlined above. In other words, summative language testing is availing itself of the most powerful tool of our generation to improve the efficiency and effectiveness of examinations. With most language learners using computers, smartphones and the like every day (and very likely using them for communication in English too), it is natural and fitting that tests should use the same medium.

However, we risk taking one step forward with technology while taking two steps back in testing language. Firstly, the advantages of CALT as it has developed are largely those contributing to greater *reliability* (for clear and simple definitions of the technical concepts 'reliability' and 'validity' see Hughes [6], chapters 4 and 5, pp.22-43). Improvements in administration, scoring, the production and publication of results, and item selection are all essentially changes that affect reliability: greater security and repeated test dates, numerous test items and more standardised marking are all huge benefits, but there is precious little in the CALT revolution that addresses issues of *validity* or language content. Indeed, CELA's computer-based tests are identical to their traditional versions, and the OTE is specifically anchored to the CEFR, making linguistic content dependent upon criteria produced over twenty years ago and largely before the 'computer age' in

language testing. There are few improvements to the validity of language testing offered by CALT so far, the only possible area for this being in CAT and here the jury is certainly still out.

Indeed, this underlines the second backward step we risk taking: the content of current CALT is hardly ever informed by the nature of the medium itself. Slavish copies of paper-based tests cannot be expected to exploit the benefits of computer technology to the full. On the contrary, the danger is that we try to force one type of language into an unnatural context. Nowhere is this more evident than in listening tests that are carbon copies of the limitations of their cd-based predecessors. Producing instruments with accompanying video (thus making supposed conversations closer to most authentic interactions with their visual elements) might allow students to assess their integrated listening and observational skills, much as they would in a real-life context.

In other words, the medium should contribute to test design and construction, not merely offer itself as another way to do the same thing. This is all the more the case, if we consider that language use is undergoing change as it operates in different contexts (Skype, weblogs, virtual meetings, tweets etc.). If we bear in mind that (especially summative) tests can be authoritative and patronising (Shohamy 2001:124) and have significant *washback* effects, it is all the more important that we ensure their validity.

References

[1] Oxford University Press. 2012. *The Oxford Test of English White Paper.* Oxford, OUP
[2] Oxford University Press. 2012. *The Oxford Test of English White Paper Summary.* Oxford, OUP
[3] Pathan, M.M. 2012. "Computer Assisted Language Testing [CALT]: Advantages, Implications and Limitations" *Researchvistas.com Vol. 1 Issue 4*, 2012
[4] Shohamy, E. *The Power of Tests.* Harlow, Longman, Pearson Education, 2001
[5] Cambridge English. *Research Notes, Issue 22,* 2005
[6] Hughes, A. *Testing for Language Teachers*. Cambridge, CUP
[7] Cambridge English. *Research Notes, Issue 51,* 2013

[1]Examples are: Cambridge CELA, who now offer their *Main Suite* examinations in an alternative computer-based form, and, most recently, OUP with their *Oxford Test of English*. See *The Oxford Test of English White Paper Summary, December 2012* for a brief description, and sample papers available online.

[2] When attempting the revised CELA Proficiency sample examination, 2013, on computer, the present author experienced considerable practical difficulty in part six (paragraphing) of the reading paper. For a discussion of the new CELA Proficiency examination, see Research Notes, 51 [7].