

# Computer-based IELTS and paper-based versions of IELTS

TONY GREEN AND LOUISE MAYCOCK, RESEARCH AND VALIDATION GROUP

## Introduction

A linear computer-based (CB) version of the IELTS test is due for launch in 2005. The CB test will, in the context of growing computer use, increase the options available to candidates and allow them every opportunity to demonstrate their language ability in a familiar medium. As the interpretation of computer-based IELTS scores must be comparable to that of paper-based (PB) test scores, it is essential that, as far as is possible, candidates obtain the same scores regardless of which version they take.

Since 2001, the Research and Validation Group has conducted a series of studies into the comparability of IELTS tests delivered by computer and on paper. Early research indicated that we could be confident that the two modes of administration do not affect levels of performance to any meaningful extent. However, the findings were muddled by a motivational effect, with candidates performing better on official than trial tests. To encourage candidates to take trial forms of the CB test, these had been offered as practice material to those preparing for a live examination. However, candidates tended not to perform as well on these trial versions (whether computer- or paper-based) as they did on the live PB versions that provided their official scores.

This report relates to the findings of the first of two large scale trials, referred to as Trial A, conducted in 2003–2004. In these studies, to overcome any effect for motivation, candidates for the official IELTS test were invited to take two test versions at a reduced price – a computer-based version and a paper-based version – but were not informed which score would be awarded as their official IELTS result.

## Previous studies of CB and PB comparability

When multiple versions or ‘forms’ of a test are used, two competing considerations come into play. It could be argued that any two test forms should be as similar as possible in order to provide directly comparable evidence of candidates’ abilities and to ensure that the scores obtained on one form are precisely comparable to the scores obtained on another. On the other hand, if the forms are to be used over a period of time, it could be argued that they should be as dissimilar as possible (within the constraints imposed by our definition of the skill being tested) so that test items do not become predictable and learners are not encouraged to focus on a narrow range of knowledge. On this

basis, Hughes (1989) argues that we should ‘sample widely and unpredictably’ from the domain of skills we are testing to avoid the harmful backwash that might result if teachers and learners can easily predict the content of the test in advance. Indeed, this would pose a threat to the interpretability of the test scores as these might come to reflect prior knowledge of the test rather than ability in the skills being tested.

Different forms of the IELTS test are constructed with these two considerations in mind. All test tasks are pre-tested and forms are constructed to be of equal difficulty (see Beeston 2000 for a description of the ESOL pretesting and item banking process). The test forms follow the same basic design template with equal numbers of texts and items on each form. However, the content of the texts involved, question types and targeted abilities may be sampled differently on each form. The introduction of a CB test raises additional questions about the comparability of test forms: Does the use of a different format affect the difficulty of test tasks? Do candidates engage the same processes when responding to CB tests as they do when responding to PB tests?

Earlier studies of IELTS PB and CB equivalence have involved investigations of the receptive skills (Listening and Reading) and Writing components. The Speaking test follows the same face-to-face format for both the CB and PB test formats and so is not affected by the CB format.

Shaw et al (2001) and Thighe et al (2001) investigated the equivalence of PB and CB forms of the Listening and Reading IELTS components. Shaw et al’s study (*ibid.*) involved 192 candidates taking a trial version of CBIELTS shortly before a different live PB version of the test which was used as the basis for their official scores. The CB tests were found to be reliable and item difficulty was highly correlated between PB and CB versions ( $r = 0.99$  for Listening,  $0.90$  for Reading). In other words, test format had little effect on the order of item difficulty. Correlations (corrected for attenuation) of  $0.83$  and  $0.90$  were found between scores on the CB and PB versions of Listening and Reading forms respectively, satisfying Popham’s (1988) criterion of  $0.8$  and suggesting that format had a minimal effect on the scores awarded. However, Shaw et al (*ibid.*) called for further investigation of the comparability of PB test forms as a point of comparison.

The Thighe et al (2001) study addressed this need. Candidates were divided into two groups: *Live* candidates comprised 231 learners preparing to take an official IELTS test at eight centres worldwide who took a trial form of either the Reading or Listening

**Table 1: Agreement rates of live and preparatory candidates for Reading and Listening (Thighe et al 2001)**

	Live candidates		Preparatory candidates	
	Reading	Listening	Reading	Listening
% agreement	30%	27%	27%	25%
% agreement to within half a band	68%	62%	61%	68%
% agreement to within a whole band	89%	89%	85%	91%

**Table 2: Agreement rates for Reading, Listening, Writing, and Overall scores in Trial A**

	Reading	Listening	Writing	Overall†
% agreement	26%	22%	53%	49%
% agreement to within half a band	72%	62%	*	95%
% agreement to within a whole band	91%	85%	92%	100%

\* Scores for Writing tests are awarded in whole band increments † Note that overall scores for the two tests (CB and PB) include a common Speaking component

component of PB IELTS two weeks before their official ‘live’ test, which was then used as a point of comparison; *Preparatory* candidates were 262 students at 13 centres who were each administered two different trial forms of either the Reading or Listening PB component with a two week interval between tests. Table 1 shows rates of agreement – the percentage of candidates obtaining identical scores, measured in half bands, on both versions of the test – between the different test forms. Half band scores used in reporting performance on the Reading and Listening components of IELTS typically represent three or four raw score points out of the 40 available for each test. For the Live candidates, who more closely represented the global IELTS candidature, there was absolute agreement (candidates obtaining identical band scores on both test forms) in 30% of cases for Reading and 27% of cases for Listening. 89% of scores fell within one band on both test occasions. The rates of agreement found between PB test versions would serve as a useful benchmark in evaluating those observed in the current study.

For IELTS Writing, the difference between the CB and PB formats is mainly in the nature of the candidate’s response. On the PB test, candidates write their responses by hand. For CB they have the option either of word-processing or hand-writing their responses. Brown (2003) investigated differences between handwritten and word-processed versions of the same IELTS Task Two essays. Legibility, judged by examiners on a five-point scale, was found to have a significant, but small, impact on scores. Handwritten versions of the same script tended to be awarded higher scores than the word-processed versions, with examiners apparently compensating for poor handwriting when making their judgements. Shaw (2003) obtained similar findings for First Certificate (FCE) scripts.

A study by Whitehead (2003) reported in *Research Notes* 10 investigated differences in the assessment of writing scripts across formats. A sample of 50 candidates’ scripts was collected from six centres which had been involved in a CBIELTS trial. Candidates had taken a trial CB version of IELTS followed soon afterwards by their live pen-and-paper IELTS; thus for each candidate a

handwritten and a computer-generated writing script was available for analysis. For Whitehead’s study, six trained and certificated IELTS examiners were recruited to mark approximately 60 scripts each; these consisted of handwritten scripts, computer-based scripts and some handwritten scripts typed up to resemble computer-based scripts. The examiners involved also completed a questionnaire addressing the assessment process and their experiences of, and attitudes to, assessing handwritten and typed scripts. Whitehead found no significant differences between scores awarded to handwritten and typed scripts. Although CB scripts yielded slightly lower scores and higher variance, Whitehead suggests that these differences could be attributable to the motivation effect described above.

Although response format seemed to have little impact on scores, Brown (2003), Shaw (2003) and Whitehead (2003) all identified differences in the way that examiners approach typed and handwritten scripts. IELTS examiners identified spelling errors, typographical errors and judgements of text length in addition to issues of legibility as areas where they would have liked further guidance when encountering typed responses. One response to this feedback from examiners has been to include a word count with all typed scripts, an innovation that was included in the current study.

## CBIELTS Trial A 2003–2004

627 candidates representing the global IELTS test-taking population took one CBIELTS Listening form and one CBIELTS Academic Reading form, alongside one of three CB Writing versions. Each candidate took the computer-based test within a week of taking a live paper-based test (involving 18 different forms of the PB test). Half of the candidates were administered the CB test first, the other half took the PB test first. Candidates could choose whether to type their answers to the CB Writing tasks or to hand-write them. All candidates took only one Speaking test, since this is the same for both the PB and CB tests. The candidates (and Writing examiners) were not aware of which form would be used to generate official scores and so can be assumed to have treated

both tests as live. Candidates were also asked to complete a questionnaire covering their ability, experience and confidence in using computers as well as their attitudes towards CBIELTS. The questionnaire was administered after the second of the two tests and results will be reported in a future issue of *Research Notes*.

Of the 627 candidates who took part in the trial, 423 provided a complete data set, including responses to two test forms and the questionnaire. Despite a slightly higher proportion of Chinese candidates in the sample compared with the live population, the sample represented a range of first languages, reasons for taking IELTS, level of education completed, gender and age groups.

## Findings

Table 2 shows rates of agreement between the band scores awarded on the CB versions with band scores awarded on the PB versions. The figures for absolute agreement are similar to, albeit slightly lower than those obtained in the earlier trials comparing PB test forms, while agreement to within half a band is slightly higher. The similarity of the results suggests that the use of a different test format (CB or PB) has very little effect on rates of agreement across forms with nearly 50% of candidates obtaining an identical band score for the test on both occasions and a further 45% obtaining a score that differed by just half a band on the nine band IELTS scale (see Overall column).

Although the results suggested that format has a minimal effect on results, some areas were identified for further investigation (Maycock 2004). Among these it was noted that, for Writing, candidates performed marginally better on the paper-based test than on the computer-based test. It was suggested that this could be due to differences in task content between versions, the effects of typing the answers, or differences in the scoring of typed and handwritten scripts.

To respond to this concern, a follow-up study was implemented to identify sources of variation in the scoring of writing scripts. The study involved 75 candidates selected to represent variety in L1 background and IELTS band score (Green 2004). Their scripts included responses to both computer- and paper-based versions of the test. All handwritten responses (all of the PB scripts and 25 of the 75 CB scripts) were transcribed into typewritten form so that differences in the quality of responses to the two exam formats could be separated from differences attributable to presentation or to response mode. Multi-faceted Rasch analysis was used to estimate the effects of test format (CB/PB) response format (handwritten/typed) and examiner harshness/leniency on test scores. The evidence from the study indicated that there was no measurable effect of response type and that the effect of test format, although significant, was minimal at 0.1 of a band.

## Conclusion

A further trial (Trial B) of additional forms of CBIELTS has just been completed and analysis is underway. The evidence gathered to date suggests that CBIELTS can be used interchangeably with PB IELTS and that candidates, given adequate computer familiarity, will perform equally well on either version of the test. However, Trial A has raised issues of scoring and the treatment of errors that will need to be addressed through examiner training and guidance. The marking process and how examiners are affected by scoring typed rather than handwritten scripts will be a continuing area of interest and will be explored further in Trial B. Initial analysis of questionnaire data suggests that candidates are generally satisfied with the CB version of the test and regard it as comparable to the PB version.

Additional questions remain regarding the processes that candidates engage in and the nature of the language elicited when taking tests with different formats. To address this, work has been commissioned by Cambridge ESOL to investigate candidate test taking processes on CB and PB tests and this is currently being undertaken by Cyril Weir, Barry O'Sullivan and colleagues at the Centre for Research in Testing, Evaluation and Curriculum in ELT at Roehampton University.

## References and further reading

- Beeston, S (2000) The UCLES EFL Item Banking System, *Research Notes*, 2, 8–10.
- Brown, A (2003) Legibility and the Rating of Second Language Writing: An Investigation of the Rating of Handwritten and Word-processed IELTS Task Two Essays, in *IELTS Research Reports Volume 4*, Canberra: IDP: IELTS Australia.
- Green, A (2004) *Comparison of Computer and Paper Based Versions of IELTS Writing: A further investigation of Trial A data*, Cambridge: Cambridge ESOL internal Validation report 585.
- Hughes, A (1989) *Testing for Language Teachers*, Cambridge, Cambridge University Press.
- Maycock, L (2004) *CBIELTS: A Report on the Findings of Trial A (Live Trial 2003/04)*, Cambridge: Cambridge ESOL internal Validation report 558.
- Popham, W J (1988) *Educational Evaluation* (2nd edition), Prentice Hall: New Jersey.
- Shaw, S (2003) Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts, *Research Notes* 11, 7–10.
- Shaw, S, Jones, N and Flux, T (2001) *CBIELTS – A comparison of computer-based and paper versions*, Cambridge: Cambridge ESOL internal Validation report 216.
- Thighe, D, Jones, N and Geranpayeh, A (2001) *IELTS PB and CB Equivalence: A Comparison of Equated Versions of the Reading and Listening Components of PB IELTS in relation to CB IELTS*, Cambridge: Cambridge ESOL internal Validation report 288.