

Computerizing an English language placement test

Glenn Fulcher

This article considers the computerization of an English language placement test for delivery on the World Wide Web (WWW). It describes a pilot study to investigate potential bias against students who lack computer familiarity or have negative attitudes towards technology, and assesses the usefulness of the test as a placement instrument by comparing the accuracy of placement with a pencil-and-paper form of the test. The article focuses upon the process of considering (and discounting) rival hypotheses to explain the meaning of test scores in the validation process.

Introduction

This paper is primarily concerned with the delivery of a language test over the Internet, using the World Wide Web. This is a computer-based test (CBT), used for placing students into 'upper intermediate' or 'advanced' classes on summer and pre-sessional courses at a UK university. It consists of 80 multiple-choice grammar items and two essays. The test is usually administered on the first day of each course, and the staff grade all tests in time to place students in classes the following morning. With as many as 120 students on a course, the task of grading the papers requires the availability of a large number of markers for a considerable period of time. Converting the multiple-choice section of the test into a CBT would remove this work from the staff, creating additional time for pedagogic activities such as class preparation and meeting students individually on the first day of the course. Essays would, however, still need to be graded by course staff.

In order to cope with large numbers of students taking tests within a short period of time it is essential that any workstation in the university can be used for testing, particularly those in larger computer laboratories. In order to avoid problems associated with testing software being loaded onto machines by technicians prior to every test session, and removed afterwards, it became clear that a platform and local area network-independent solution was required. An Internet-delivered test can be kept permanently on a server, with access permissions switched on just before testing begins, and switched off as soon as it is over. The students can be invigilated in multiple computer laboratories at any point in the university, and they can access the test by a Uniform Resource Locator (URL), which is announced at the beginning of the test. The software required is any standard Internet browser, which makes test delivery platform-independent. Internet delivery therefore frees the administration from the requirement to use a particular site on the campus, and it does not matter whether the invigilators and the

course co-ordinators are able to book PC or Macintosh laboratories.

Research questions

The introduction of a new medium for assessment requires considerable thought and research. Although the motivation for a change in the delivery of the multiple-choice section of the test is to create time for pedagogic reasons, it is vital that the information gained from test scores is appropriate for the decisions that will be based on them. This is a matter of equity and fairness in testing practice.

The project described in this article, therefore, was motivated by an awareness that educators had to ensure the introduction of a CBT delivered on the Internet, using a standard Internet browser, in such a way that it would not disadvantage any students. It was also essential for the CBT to provide accurate information upon which appropriate placement decisions could be made. Specifically, we wished to know:

- (a) Is the objective part of the placement test sufficiently reliable for its purpose?
- (b) How well are the CBT and pencil-and-paper versions of the test correlated?
- (c) Is the CBT better or worse at placing students into appropriate classes?
- (d) Are scores significantly affected by familiarity with computers, attitudes to computer delivery of tests, or test-taker background?

The questions were addressed using a sample of 57 students who took both the paper-and-pencil and the computer-based forms of the test in 1997. Whilst this is not a large sample, it is nevertheless adequate to investigate the first three questions with confidence, and to test for main effects, if not interactions, on the final question.

Equity in computer-based testing

The illusory equivalence of forms

Computer-based tests are subject to the same standards of reliability and validity that would be expected of any other test. However, in the case of computer-based tests, certain critical issues of equity have been raised. The classic statement of these concerns may be summarized from the *Guidelines for Computer Based Tests and Interpretations* (APA: 1986):

Test developers should demonstrate that paper-and-pencil and computer-based versions of a test are equivalent forms.

The two forms of the test should rank test-takers in approximately the same way.

The means and standard deviations of the two forms should be approximately the same.

These guidelines give an indication of what is to be expected of a computer-based test that has been converted from a test that already exists in paper-and-pencil format. The criteria laid down by the APA relate to the strong possibility that in the translation of the test from one medium to another, the new medium may change the nature of the construct underlying the test, or alter the scale. These are serious issues. However, the requirement that forms are equivalent is somewhat

restrictive. It is, for example, quite possible to imagine a situation in which a CBT form provides better placement information than its pencil-and-paper form. If the scale has been altered, the users of test scores must be certain that with the introduction of the CBT the new score meaning is understood and interpreted appropriately.

A great deal of useful work has been conducted into the equivalence of paper-and-pencil with computer-based forms of tests, most of which provide evidence that achieving equivalence is difficult. Bunderson *et al.* (1989) provide an overview of studies of test equivalence to the end of the last decade which shows that three studies had recorded a higher score on computer-based tests, 13 had shown higher scores on the paper-and-pencil test, and 11 had shown no difference in scores across forms. They concluded that lack of familiarity with computers could be assumed to be a major factor in achieving lower scores on computer-based tests, but that scores on paper-and-pencil tests were lower for younger test-takers who had not been familiarized with the method of completing the answer booklet. The issue of familiarity is not new in language testing. It has always been accepted that test-takers should be familiar with the item types and mode of test delivery before taking a test to ensure that these factors would not be confounding variables in score interpretation. On a similar point, Russell and Haney (1997) recently suggested that it is becoming increasingly unfair to test writing ability by paper-and-pencil in situations where learners have become used to composing directly on word processors. In this case, one would expect the computer-delivered version of a writing test to result in higher scores simply because of familiarity. Yet the higher score would more accurately reflect the ability to be tested. Familiarity with test format and item types is essential for all testing, and however much it has become associated with it in recent literature it is not specific to computerized testing.

The most recent meta-analysis of equivalence of paper-and-pencil with computer-based forms of tests has been conducted by Mead and Drasgow (1993). This study is concerned, as most other studies have been, with the method of delivery—whether by paper-and-pencil or computer. It also includes two other variables: conventional vs. adaptive tests and power vs. speeded tests. The difference between conventional and adaptive tests is not relevant to the project discussed in this paper, but speededness should be considered. It seems likely that the speededness of the test may affect scores across forms. On the computer screen, the fact that scrolling has to be operated by the test-taker could result in a score reduction on the computer-based form. Mead and Drasgow (1993) report that in the meta-analysis of 28 studies, the computer tests were slightly harder than their paper-and-pencil counterparts, but that the only variable to significantly affect scores was speededness. Whilst the correlation between tests was on average 0.91 for different administration modes, the average cross-mode correlation for speeded tests was 0.72. One possible explanation for this finding is the differences in motor skills required of paper-and-pencil compared to

computer-based response techniques, when working under severe time limits. However, timed-power tests show no significant influence of medium on test scores. Mead and Drasgow (1993: 456) conservatively state that although 'a computerized version of a timed power test can be constructed to measure the same trait as a corresponding paper-and-pencil form' it should not be taken for granted for any computerized test. This should be a warning that future computer-based tests should be submitted to the same rigorous scrutiny as previous computerized tests, especially if they are speeded.

This review indicates that any attempt to achieve equivalence of forms is likely to fail. Scores will vary between the pencil-and-paper and CBT forms of a test. However, it is still incumbent upon an institution to ensure that significant score variations associated with a CBT do not introduce a bias into the results. As we have seen, this is as relevant to the investigation of pencil-and-paper forms as it is to the CBT.

Further equity issues

Although the issue of equivalence of forms has dominated the discussion of computer-based testing, this is not the only equity concern. Other equity issues relate directly to:

Previous experience of using computers. Factors in this category include the familiarity of test-takers with the computer itself, the frequency with which a computer is used (if at all), and familiarity with the manipulation of a mouse with two buttons. If the test is delivered over the Internet using a standard browser such as Netscape or Internet Explorer, familiarity with the WWW (perhaps also including frequency of e-mail use) should be taken into account.

Attitudes of test-takers to computers, the software being used (the degree of user-friendliness, for example), and the WWW in general, also need to be investigated. It is at least feasible to suggest that negative attitudes to the medium of delivery could have an impact upon test scores, and this may be more likely among those with little or no experience of using computers or the Internet.

Finally, the background of the test-taker may be relevant to the validity of the test score. Background factors worthy of investigation would seem to be age, gender, primary language background (L1), and level of education or subject specialism in the case of applicants for university places.

The largest move to computer-based testing was the introduction of the computer-based TOEFL in 1998. Educational Testing Service (ETS) has conducted a significant amount of research on the TOEFL takers' access to and experience with computers, in their attempt to design a computer-based TOEFL that is minimally dependent upon previous experience. In the case of ETS, this has involved the development of a tutorial package that test-takers do prior to taking the TOEFL.

Taylor, Jamieson, Eignor, and Kirsch (1997; 1998) investigated computer familiarity in the TOEFL test-taking population and its effect

on test scores, including gender and L1 as variables. Using analysis of covariance (ANCOVA), Taylor *et al.* argue that significant statistical relationships between some of the sub-tests of the TOEFL and computer familiarity were so small as to be of no practical importance, and that the same picture emerged for gender and geographical location. In total, the scores of around 20% of TOEFL test-takers may be affected by the computer-based form, although it is assumed that the tutorial package will further minimize this. Despite the claims of Taylor *et al.*, introducing the computer-based test will have a significant impact on at least one fifth of test-takers. This is not an insignificant number, and so further research is needed. It is also interesting to note that Educational Testing Services decided not to introduce the computer-based TOEFL into East Asia as planned (Educational Testing Services 1998).

Method The reliability of the 80-item objective component of the pre-sessional placement test was investigated using a sample of 57 students enrolled on pre-sessional courses during 1997. The test is entirely multiple-choice, and has been used in the pre-sessional programme at the university for five years in its current format. Although the test has been little studied, the teachers who mark it and use it to place students in class have consistently reported that they are happy with the placement decisions taken on the basis of the scores, in conjunction with the essay results.

The test was computerized, using QM Web (Roberts 1995; see also Fulcher 1997), and loaded onto a web server. Authorizing access to the test, downloading test scores on the server, and compiling the test results, was undertaken remotely from a personal computer.

Fifty-seven students enrolling for courses in January 1997 took the test in a standard university computing laboratory using the Netscape 3.0 Internet browser, and its paper-and-pencil form. The test is a timed-power test, and was designed to take 45 minutes in total. The computing laboratory was invigilated by two course tutors, who helped students log on to the test, and ensured that each test-taker submitted the answers for scoring on completion of the test. The computer-based test was taken within two weeks of the paper-and-pencil test. It is acknowledged that some order effect is inevitable. Nevertheless, this pragmatic decision not to randomize the order of test-taking was taken because students must be placed into classes on arrival, and it would be unfair to use the CBT to do this itself before its properties could be established. However, this is not a confounding factor in the design of the study, since the matter of strict equivalence is not an issue only of the relationship of the two tests and the power of student placement, or classification. The only impact of the ordered administration is on the increase in mean scores on the CBT.

After taking the CBT, data was collected from each student on computer familiarity. This recorded frequency of computer use, familiarity with the mouse, familiarity with the WWW, and frequency of e-mail use. Attitudes towards taking tests on the Internet included a

preference for paper-and-pencil format, a preference for Internet format, and a student-estimated likelihood of getting a higher grade on one test or the other. Background information on the test-taker included age, gender, Primary Language Background, and subject area specialism (the subject students intend to study at university).

The data were analysed to compare the test scores for the sample of students across the two test formats. Scores on the computer-based test were investigated using ANCOVA for the effect of computer-familiarity, attitudes, and the test-taker's background. In this process the paper-and-pencil based test was used as a covariate to take ability into account.

Results and discussion
Reliability

Test reliability was estimated using Cronbach's alpha. Reliability for the paper-and-pencil test was estimated at 0.90, while the reliability of the computer-based test was 0.95. This indicates that further investigation into variation in test scores should not be confounded by random error.

Relationship of the paper-and-pencil form to the WWW form

First, descriptive statistics were calculated for the two forms of the test (see Table 1). The difference between the mean of the WWW form (CBT) and the paper-and-pencil form (TEST) is significant, as shown in Table 2, the paired samples t-test. Table 3 presents the correlation between the two forms, which is respectably high for this type of study. It is accepted that the increase in mean score on the CBT is due in large part to an order effect, but this in itself is not enough to account for the possible variation in scores as indicated by the standard deviation of the difference of means.

Table 1. Descriptive statistics (scores as percentages)

	N	Minimum	Maximum	Mean	Standard Deviation
Paper-and-pencil	57	29	78	56.58	13.76
CBT	57	19	80	62.23	14.16

Table 2. Paired samples t-test for the two forms

	Mean	S.D.	Paired Differences		t	df	Sig. (2-tailed)	
			Standard Error of Means	95% Confidence Interval of the Difference				
				Lower				Upper
Pair CBT-Test	5.65	8.29	1.10	3.45 7.85	5.14	56	.00	

Table 3. Correlation between the two forms

	TEST	CBT
Pearson Correlation		.82*

*Significant at the 0.01 level (2 tailed test)

The standard deviation of the difference of means (see Table 2) is 8.29. This means that for 95% of the sample, the score on the CBT could be anywhere from 10.6 lower to 21.9 higher than their score on the paper-and-pencil form of the test. This is an indication that a correlation of 0.82 is not high enough to ensure that the CBT is an adequate replacement for the paper-and-pencil version of the test. However, as we have argued above, this does not mean to say that the CBT is not a better placement

instrument than the pencil-and-paper form of the test. We return to this issue below.

Bias In Tables 4 to 6, the areas of computer familiarity, attitudes towards taking tests on the Internet, and towards the test-taker's background were investigated. The CBT was the dependent variable in each case, and the paper-and-pencil test (labelled "TEST") was used as a covariate to take ability into account. In each case Analysis of Covariance (ANCOVA) is used. It should be stressed that with 57 cases these results should be treated with some caution; whilst it is possible to test for main effects, interaction effects (if present) could not be detected without a much larger dataset.

The independent variables of frequency of computer use, familiarity with the mouse, and frequency of use for the WWW and e-mail (listed in the left hand column) were measured on a five-point Likert scale. Table 4 shows that there are no significant main effects on the CBT scores.

Table 4. Computer familiarity

Source	df	F	Sig.
TEST	1	40.00	.00
FREQUENCY	4	1.46	.26
MOUSE	4	1.42	.28
WWW	4	.17	.95
EMAIL	4	.86	.51

In measuring attitudes towards testing format, test-takers were asked to respond to two five-point Likert items, one of which asked if they liked taking paper-and-pencil tests (LIKETEST) and one which asked if they liked taking the Internet-based test (LIKECBT). Next, they were asked to say, with only two options, on which test they thought they would get the highest score (CBTBETTER), and finally, if they were given a choice of paper-and-pencil or a computer-based test, which one would they choose. Again, there was no significant impact of attitudes on the computer-based test scores, as demonstrated in Table 5.

Table 5. Attitudes towards taking tests on the Internet

Source	df	F	Sig.
TEST	1	19.25	.00
LIKE CBT	3	0.37	.78
LIKETEST	4	1.19	.35
CBTBETTER	3	2.00	.15
CHOICE	2	0.25	.78

Finally, the background of the test-taker was investigated. The Primary Language Background (L1) was classified only as Indo-European or non-Indo-European. As all the test-takers were applying to attend courses at UK universities, their subject areas were classified into one of three categories: science, engineering, or humanities and social sciences. The results are presented in Table 6.

Table 6. Background of the test-taker

Source	df	F	Sig.
TEST	1	76.63	.00
AGE	16	1.82	.126
GENDER	1	0.00	.98
L1	1	6.56	.02
SUBJECT	2	1.12	.35

The only significant effect discovered was that of Primary Language Background (L1). The mean score of students speaking Indo-European languages on the computer-based test was 64.18, with a standard deviation of 14.13. Speakers of non-Indo-European languages, on the other hand, had a mean of 61.40, and a standard deviation of 14.31. However, on the paper-and-pencil test, there was no significant difference between these two groups. There does, therefore, appear to be a possibility that some learners (mostly from Japan and Korea, in this sample) may be placed into a lower group on the CBT.

Bias or better placement? Examining competing hypotheses in validity studies

One possible explanation for this result may lie in the principle of 'uncertainty avoidance'. In cross-cultural studies conducted by Hofstede (1983, 1984; see also discussion in Riley 1988) it has been demonstrated that Japanese subjects suffer a greater degree of stress than other nationalities when asked to do unfamiliar tasks. Hofstede was conducting research for IBM into the computerization of business tasks in industry. It was discovered that workers who were familiar with using computers in the workplace, and familiar with carrying out certain tasks, suffered from significantly increased stress levels when asked to do the task on the computer. The combination produced an uncertainty that in turn created more stress, and an increase in the likelihood of making mistakes. In the case of computerized tests, a test-taker may be familiar with taking tests of various types, and may be familiar with computers. But if they have never taken a test on a computer, the fear of 'doing old things in new ways' becomes an important factor that can affect test scores. This particular explanation would go a long way to accounting for the significant findings presented in Table 6, whilst accounting for non-significant results for all other factors that could account for variability in test scores.

One key element in test validation is the provision of theories that best account for the empirical data (Messick: 1989). Although variability in scores across test formats are normally accounted for by theories like the one presented above, it should not be forgotten that differences between groups of students on test scores may represent differences in ability, rather than bias. This could come about because of the shared learning experiences of certain groups of students, or the distance between their L1 and the target language. It is therefore important to investigate the extent to which the CBT is better at placing students into coherent teaching groups—arguably the most important validity criterion for the evaluation of a placement test.

It will be recalled that when the pencil-and-paper test is used, placement decisions are only made after two pieces of written work have been graded. Teachers then meet to consider all grades and assign students to one of two initial placement levels. It is assumed that this longer process, including evidence from the writing tasks, is more accurate than relying on the results of the objective test alone. Inter-rater reliability for assessment of the writing is calculated at 0.87. The question that needs to be asked is whether the CBT is better than the pencil-and-paper form of the test on its own at predicting the final placement decision. If it is, the CBT would provide better quality information to teachers making placement decisions than the pencil-and-paper form of the test. (Note that the writing test would still be used: the question posed here is related to the quality of information provided by one part of the test. Improving this increases the utility of the test score as a whole.) In Table 7, we can see the results of a one-way ANOVA study to investigate this question. It can be seen that the mean scores of the final placement groups are significantly different on the CBT, but not significantly different on the pencil-and-paper test. In other words, the CBT predicts final decisions more accurately than the pencil-and-paper test.

Table 7. Discrimination between two groups on the CBT and pencil-and-paper forms

		Sum of squares	df	Mean square	F	P
CBT	Between groups	1,092,747	1	1,092,747	5,929	.018
	Within groups	10,137,289	55	184,314		
	Total	11,230,035	56			
TEST	Between groups	176,932	1	176,932	930	.339
	Within groups	10,460,962	55	190,199		
	Total	10,637,895	56			

Comparing the means of Group 1 (advanced) with Group 2 (upper-intermediate), we can see the greater discriminatory power of the CBT in Table 8.

Table 8. Means of groups by test form

Group	Advanced			Upper-intermediate	
	Mean	Sd.	Std. error	Mean	Sd.
CBT	71.18	8.89	2.68	60.09	14.41
TEST	60.18	7.45	2.25	55.72	14.84

We have presented two cases. The first suggests that the CBT scores are contaminated by an L1 background factor, making placement decisions suspect and unfair to students from East Asia. The second case suggests that the CBT is more sensitive to variation in language ability (likely to be related to L1 background) and contributes more information than the pencil-and-paper test to a fair and valid placement decision. In this second case, there is an implication that it is the mode of delivery of the pencil-and-paper form that interferes with the discriminatory power of the test items.

The interpretation of the data that is preferred by the institution will have an impact upon the immediate future learning environment of the student. It is incumbent upon test developers and users within institutions to be aware of the rival interpretation of test scores, and to be aware of the impact of accepting one hypothesis over another too readily or, worse, without considering possible rival hypotheses at all.

Conclusions In answer to the research questions posed, it has been discovered that the CBT is sufficiently reliable for its purpose, and that the two forms are correlated—but not highly enough for accurate prediction of a score on one form of the test from a score on the other form. It has been discovered that the CBT provides better information than the pencil-and-paper test in placing students into one of two groups, but that there may be some bias against students with certain L1 backgrounds.

The question that remains for this institution is whether to implement the CBT (alongside the existing writing tests) on the basis of this evidence, or to conduct further research that would provide further evidence in support of one of the two hypotheses regarding score meaning outlined above. In practice, the pedagogic and practical advantages to be gained from the CBT are too valuable to waste. As the CBT appears to contribute more to the decision-making process than the pencil-and-paper test, it is therefore likely that the test will be put into operational use. Any bias present in the test would mean that the institution would have a small number of false negatives (students placed into a lower group who should be placed in a higher group), but these individuals are usually identified within a short period of time, and moved accordingly by the teachers. These placement tests are not sufficiently 'high stakes' to cause undue worry.

The gains therefore outweigh the seriousness of the possible presence of bias in the CBT. Teachers will be made aware of this possibility, and additional attention will be paid to students from these L1 backgrounds in the grading of the essay component of the test, and during the introductory classes.

This article has looked at the issues facing a particular institution over the introduction of a computer-based placement test as a replacement for its pencil-and-paper multiple-choice test. In the process, it has highlighted the importance for any institution of considering rival hypotheses of score meaning when changes in assessment practice are being reviewed.

Received November 1998

References

- APA.** 1986. *Guidelines for Computer Based Tests and Interpretations*. Washington DC: American Psychological Association.
- Bunderson, C. V., D. I. Inouye, and J. B. Olsen.** 1989. 'The four generations of computerized educational measurement' in R. L. Linn (ed.). *Educational Measurement* (3rd edn.). Washington, D.C.: American Council on Education. 367-407.
- Educational Testing Services** 1998. *TOEFL 1998 Products and Services Catalogue*. New Jersey: ETS.
- Fulcher, G.** 1997. 'QM Web: A World Wide Web test delivery program'. *Language Testing Update* 21: 45-50.
- Hofstede, G.** 1983. 'The Cultural Relativity of Organizational Practices and Theories'. *Journal of International Business Studies* 83: 75-89.
- Hofstede, G.** 1984. *Culture's Consequences: International Differences in Work Related Values*. Vol. 5: Cross-Cultural Research and Methodology Series. Beverly Hills: Sage.
- Mead, A. D. and F. Drasgow.** 1993. 'Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis'. *Psychological Bulletin*, 114/3: 449-58.
- Messick, S.** 1989. 'Validity' in R. L. Linn (ed.). *Educational Measurement*. New York: Macmillan: 13-104.
- Riley, P.** 1988. 'The Ethnography of Autonomy' in A. Brooks and P. Grundy (eds.). *Individualization and Autonomy in Language Learning*. ELT Documents 131: 12-34. London: Modern English Publications in Association with the British Council.
- Roberts, P.** 1995. *QM Web: Tests and Surveys on the Web*. London: Questionmark Computing.
- Russell, M. and W. Haney.** 1997. 'Testing Writing on Computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil'. *Education Policy Analysis Archives* 5/3. <http://olam.ed.asu.edu/epaa/v5n3.html>
- Taylor, C., J. J. Jamieson, D. Eignor, and I. Kirsch.** 1997. 'Measuring the Effects of Computer Familiarity on Computer-based Language Tasks'. Paper presented at the Language Testing Research Colloquium, Orlando, Florida.
- Taylor, C., J. J. Jamieson, D. Eignor, and I. Kirsch.** 1998. *The Relationship Between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks*. New Jersey, Educational Testing Service: TOEFL Research Report 61.

The author

Glenn Fulcher is currently Director of the English Language Institute at the University of Surrey. He has worked in EFL abroad and in the United Kingdom since 1982, gaining his MA in 1985 and his PhD in language testing from the University of Lancaster in 1993.

E-mail: g.fulcher@surrey.ac.uk

Resources in Language Testing WWW page: <http://www.surrey.ac.uk/ELI/tr.html>