An English language placement test: issues in reliability and validity

Glenn Fulcher *English Language Institute, University of Surrey*

This report describes a reliability and validity study of the placement test which is used at the University of Surrey as a means of identifying students who may require English language support whilst studying at undergraduate or postgraduate levels. The English Language Institute is charged with testing all incoming students, irrespective of their primary language or subject specialization. Fair and accurate assessment of student abilities, and referring individuals to appropriate language support courses (the in-sessional programme), is an essential support service to all academic departments. The goal of placement testing is to reduce to an absolute minimum the number of students who may face problems or even fail their academic degrees because of poor language ability or study skills. This study looks at the administrative and logistic constraints upon what can be done, and assesses the usefulness of the placement test developed within this context.

I Introduction

It has recently been noted that although placement testing is probably one of the most widespread uses of tests within institutions, there is relatively little research literature relating to the reliability and validity of such measures (Wall, Clapham and Alderson, 1994). Publications which deal with placement tests frequently provide qualitative assessments of instruments (Goodbody, 1993), or are concerned with the placement of linguistic minority students in programmes which are not related to language teaching (Schmitz and DelMas, 1990; Truman, 1992). Wall, Clapham and Alderson (1994) offer one of the few empirical studies of a placement test which is designed to screen students entering a British university for deficiencies in English language skills, which might impede their progress in undergraduate or postgraduate studies. They investigate face validity (student perceptions of the test), content validity (asking whether tutors thought that test content represented programme content), construct validity (through a correlational study), concurrent validity (with self-assessment, and the assessment of tutors) and reliability. The main problem they discovered in attempting to validate the University of Lancaster placement test was in finding appropriate external criteria to conduct a concurrent study.

Language Testing 1997 14 (2) 113–138 0265-5322(97)LT127OA

The methodology used to ascertain the reliability and validity of the test in this study is similar to that of the Lancaster study in many ways, and is described in detail below. However, this study does attempt to exand on their methodology for the evaluation of placement tests within a university context in four areas. These are:

- the use of pooled judgements in establishing cut scores for placement;
- the use of additional statistical means to analyse test data;
- the consideration of the need to develop parallel forms; and
- the use of student questionnaires to investigate face validity.

II Background

The English Language Institute (ELI) has a number of roles within the University of Surrey, ranging from its support function in providing English language and study skills courses for students at the university, English for Academic Purposes (EAP) for overseas students who intend to study in an English-medium tertiary institution, and an MA in linguistics (TESOL) for teachers of English, in distance-learning mode. As part of its service function related to the in-sessional programme, the ELI is charged by the university to administer an English language test to all students entering the university on taught courses (undergraduate and postgraduate, English as primary language speakers, and speakers of other languages). The purpose of the test is to identify students whose lack of language skills or ability to communicate may cause problems in their academic work with their departments.

A major constraint in assessing students is that, together with organization (handing out and collecting papers, giving verbal instructions), the ELI must complete testing within one hour for each test administration. This means that the test cannot exceed 45 minutes, and must therefore be restricted in its length and content. Secondly, all scripts must be processed within five days of the day of the test, and results disseminated. In 1994, 1619 students took the test. The number of markers available, and the number of scripts which each can process in this time frame, also dictates that much of the paper be objectively scored. However, no language test can use only one response format, or sample only one construct, if it is to be considered valid on even prima-facie grounds (APA, 1985: 73; 75). For this reason, a writing component is included.

The placement test acts as a screening device to reduce the number of students who attend an oral interview. It is those who are invited to come to the ELI for an oral interview who have the greater probability of requiring English language support. These students are

'referred' for oral testing. A small number of students are informed after the interview that they do not need to attend the in-sessional courses. It is acknowledged (see section VI below) that all tests contain error, and it is during the interview procedure that we attempt to identify students who have been misclassified by the test. Most students are referred to a programme, although attendance is optional. Lists of all referred students are sent to heads of department, and later in the year heads are also sent an update of student attendance at courses, and a qualitative assessment of individual student progress for those who did attend.

III Test design

1 Test format

Prior to October 1994, the test contained one piece of writing, and was marked subjectively. It became clear that the ELI did not know whether this was a reliable or valid placement instrument, but it was considered inappropriate on other grounds. A single essay title may be biased in favour of some students and against others, and any single task of this nature is unlikely to elicit an adequate sample upon which decisions can be made (APA, 1985: 75; Upshur, 1971: 47; Van Weeren, 1981: 57; Shohamy, Reves and Bejerano, 1986; Shohamy, 1983; 1988; 1990). A new test was devised, with the following format:

Section 1: Essay 1: descriptive (no choice of essay title).
Essay 2: argumentative (one title to be selected from three options).
Section 2: Structure of English (10 items).
Section 3: Reading comprehension (8 items).

Essay titles for Section 1 were screened carefully to avoid bias in favour of, or against, students from any particular discipline, or culture. Essay 1 is of a general nature, whilst the titles for Essay 2 are an attempt to reflect the interests of the three faculties within the university: Human Studies, Engineering and Science. In Section 2, item content was selected for known difficulty in the in-sessional programme grammar revision course. The context of the items is general university life, on the grounds that subject-specific contexts may, once again, disadvantage some subsections of the test-taking population. In Section 3, six texts were selected, for their general academic interest. Of these, three were drawn from the humanities, and three from the sciences. However, the journals from which the texts were taken

are 'popular' in that the passages are written for the interested (intelligent) layperson, not for the specialist.

2 Pilot study

The new format was piloted during the summer of 1994, using 67 students attending the ELI Summer School English for Academic Purposes (EAP) programme. All items in Sections 2 and 3 of the test were studied using classical item analysis. Only items with a facility index between 0.3 and 0.8 were retained, and items with a discrimination index of less than 0.3 were either abandoned, or rewritten and piloted for a second time. The point biserial correlation for all remaining items was above 0.4. Four pilot versions of the test were originally written, and 75% of all questions discarded, leaving one operational test with items drawn from all four pilot tests. This is a reminder that, no matter how experienced one may be in test development, there is always a need for pretesting all items before tests become operational, and decisions are taken on the basis of test results.

In Section 1, writing samples were collected, graded by six tutors, and features of performance at each level of a rating scale established. Prototypical samples from each band level were then used in the training of raters prior to the operationalization of the test in October 1994.

IV Method

The final version of the test was used operationally with the university's entire intake on taught (undergraduate or postgraduate) courses in October 1994. The main studies were all carried out using this population.

1 Reliability

In order to establish estimates of the reliability and validity of the placement test, a number of approaches were taken. In the investigation of reliability, correlation coefficients, means and standard deviations (inter- and intrarater reliability), were established for rating patterns on Section 1 of the test. For Sections 2 and 3, it was decided to fit a logistic model. The reason for this decision was essentially because of the need, within a university setting, to have multiple forms of a test, and for the interpretation of scores on these forms to be comparable.

This implies that forms must be equated in some way. Using a logistic model, it is possible to create multiple parallel forms, using 'anchor' items from previous form(s) which can be used to calibrate

new items in later forms, even if this is difficult for short tests. In the first attempt to fit a Rasch model to the data, it became clear that there were two distinct populations within the total test-taking population. These distinct populations are those whose primary language is English, and those whose primary language is other than English. Consider Table 1, which is a simple X^2 test of significance of primary language classification and referrals, on the basis of the test scores (for a discussion of cut scores for referrals, see Section VI below). Using Yates' correction for a 2×2 grid, $X^2 = 516.61$, a highly significant result, indicating that these are certainly separate populations. It is not possible, unfortunately, to isolate differences between referrals or test scores between learners whose primary language is not English, because of the small n size of many of the primary languages represented by the 614 overseas students tested.

The question which arose, therefore, was on which population to standardize the objective components of the test. It was finally decided to standardize on the group which did not have English as a primary language, as this is the population which is more likely to require English language support. Those speakers of English as a primary language (73 in this sample) whose score profiles are similar to those of the referrals of the norming population will still be identified by the test, and remedial action can be taken.

The Rasch model allows the test developer to calibrate both item difficulty and learner ability to the same scale (Crocker and Algina, 1986: 340–41), measured in logits. This was done using the program RASCAL (Assessment Systems Corporation, 1994) for Sections 2 and 3 of the test separately. This was done as there was no theoretical reason to suspect that the combined scores of Section 2 and 3 would represent a unidimensional scale. The disadvantage of this approach, however, is that Section 2 consists of only ten items, and Section 3 of eight items. Although $n = 614$, the low number of items inevitably reduces test reliability. This could not be avoided, because of the time constraints as described above.

Table 1 X^2 table to compare primary and nonprimary English language speakers to test whether they belong to the same test-taking population

	Not referred	Referred	Total
Non-English primary language	252	362	614
English primary language	932	73	1005
Total	1184	435	1619

2 Validity

a Correlation and principal components analysis: Construct validity was assessed using correlation, and a principle components analysis. Each of the sections of the test were designed to measure different aspects of English language proficiency, and so the factorial structure of the test is an issue.

b Analysis of cut scores across referred and nonreferred students: Cut scores for the test as a whole, and each section of the test, were established after the pilot study. Scores were considered in relation to lecturers' assessments of whether students with certain profiles were ready for study in a tertiary-medium institution. Cut scores were established using this 'pooled judgements' technique (Popham, 1978: 165). Groups of essays at a range of scores were presented randomly to lecturers who were asked to decide whether this was an acceptable piece of work for undergraduate or postgraduate work in the university. Discussion was acceptable, and the cut point for each essay established at the point where the highest agreement in making dichotomous judgements was reached. However, the results require empirical investigation to ensure that the cut scores are in fact leading to appropriate decisions on whether to refer students to language support programmes.

c Concurrent validity: Thirty-three students from the overseas student population taking the test had recently taken the Test of English as a Foreign Language (TOEFL). Although this number is small, it is nevertheless possible to begin to investigate the relationship between this test and a university placement test. As more data are gathered in future years, it should be possible to establish stable concurrent validity statistics with a number of major tests which are currently used for entrance purposes.

d Content validity: Content validity was investigated by requesting three subject specialists to comment on questions set in the placement test. One was drawn from each of the three faculties within the university: Human Studies, Engineering and Physics. Only one informant suggested that the test was not content valid in one particular field.

e Feedback from students: It is clearly important to obtain qualitative feedback from students on the operation of the test. All tests have consequences for the test-takers and for the institutions which base decisions on scores. If the test is not perceived to be fair by the test-takers and score users, the role of the placement test within the

Table 2 Inter-rater reliability – Pearson product moment correlations

	Rater 1	Rater 2	Rater 3
Rater 2	0.92		
Rater 3	0.87	0.94	
Rater 4	0.75	0.83	0.93

institution is compromised. Including a qualitative study of this nature therefore relates not to the technical qualities of a testing instrument, or to accurate decision-making, but to one aspect of the social consequences of testing for the institution.

V Reliability

Reliability of subtests was initially calculated during the pilot study, and recalculated during the first operational testing. The following figures relate to the operational version of the test, with $n = 614$.

1 Section 1: writing

To establish the reliability of the assessment of writing samples, 20 essays were selected from the population, and were marked by four tutors. After a period of two weeks, the tutors were then asked to remark a subset of six samples. This allows the calculation of inter- and intrarater reliability. Table 2 shows the results of the inter-rater reliability study. It can be seen that agreement between raters is well within acceptable reliability ranges for this type of test, with an average Pearson product moment correlation of 0.87. The average band awarded was 5.46, with a standard deviation of 1.21. Table 3 shows that the four raters did not differ significantly from this in their individual grade profiles.

In the intrarater reliability study, the average reliability coefficient was 0.69, somewhat lower than the inter-rater coefficients, but still not so low as to cause undue worry. This figure indicated a need for further rater training before the second operational testing session, in

Table 3 Inter-rater reliability – Variation in means and standard deviations

	Rater 1	Rater 2	Rater 3	Rater 4	Average
Mean	5.50	5.33	5.50	5.50	5.46
SD	1.00	1.03	1.64	1.64	1.21

October 1995. The intrarater reliability coefficients were: rater 1: 0.83; rater 2: 0.57; rater 3: 0.68; and rater 4: 0.70 (Pearson product moment correlations).

2 Section 2: structure

Rasch scaling was conducted using RASCAL (Assessment Systems Corporation, 1994). Table 4 shows the questions in order of difficulty, from the easiest to the most difficult, in logits, together with the standard error associated with the difficulty estimate, and the X^2 fit statistic.

From Table 4 it can be seen that there are three misfitting items. That is, they do not meet the criterion of unidimensionality in this subtest, and must therefore be removed, and replaced by other items in this form of the test. It is instructive to return to misfitting items to attempt to provide a linguistic rationale for why they misfit. In this case, it is particularly enlightening to look at item 5, because of the very large misfit statistic. Item 5 is:

She's always _____ other students on her course, even though she is very busy herself.

a. help
b. helped
c. used to help
d. helping

With hindsight, it appears obvious that there are two possible keys to this item, but this was not spotted during pretesting and item revision. Such mistakes highlight the importance for thorough pretesting and *post hoc* analysis of all test items, on tests where important

Table 4 Rasch analysis of Section 2

Item	Difficulty	Standard error	X^2
7	–1.370	0.129	7.534
2	–1.085	0.120	10.279
1	–0.770	0.111	5.137
6	–0.668	0.109	9.298
8	–0.499	0.105	14.466 **misfit**
10	–0.085	0.099	22.565 **misfit**
3	0.128	0.096	9.597
9	0.598	0.093	8.055
4	0.745	0.093	8.055
5	3.006	0.124	45.036 **misfit**

decisions are being made. It cannot be emphasized enough that however experienced an item writer or test designer someone may be, it is not enough to rely on 'eyeballing' tests or test items.

With the test centred on item difficulty, mean item difficulty was 0.00, and the standard deviation of difficulty 1.26. Average ability was 0.89, with a standard deviation of 1.15. The test characteristic curve for Section 2 is presented in Figure 1. This plots estimated proportion of test items correct as a function of the ability of the student on the latent trait (structure of English), and may be interpreted as a nonlinear regression curve for relating raw scores to the latent trait.

When using a Rasch model, it is possible to look at reliability in two ways. First, we may calculate a reliability coefficient which is the equivalent of KR20 (an estimate of internal consistency) used in classical test analysis and, secondly, we may talk about test information. The reliability coefficient may be understood as the degree to which the test characteristic curve in Figure 1 may be relied upon as a translation of raw scores to latent trait scores. This section of the test contains only 10 items, a decision taken because of time constraints as discussed above. As reliability is related to test length, although it was hoped that reliability coefficients would be adequate, they were not expected to be very high. The reliability coefficient for Section 2 was 0.63, which is below what would be required for a high-stakes test. However, this may be reasonable for a placement test of this size.

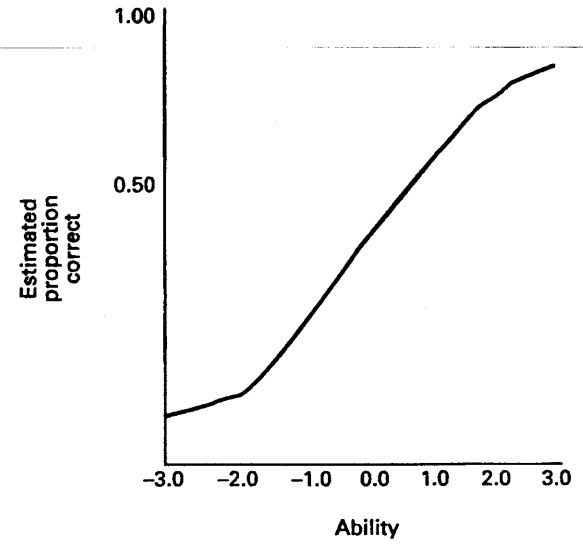

Figure 1 Test characteristic curve for Section 2 (structure)

The amount of information which a test provides differs according to region on the latent trait scale. Where the test characteristic curve is steepest, there is more discrimination amongst test-takers, and hence the test is providing greater information. Test information for Section 2 of this test is presented in Figure 2. Figure 2 indicates that Section 2 of the test is providing the most information from −1.0 to 0.5 on the latent trait scale, which is precisely what was required of this placement test. It is not necessary to discriminate finely between students at the higher end of the latent continuum. Similarly, below a certain ability level, fine discrimination is not necessary. What is crucial in this kind of testing is providing the maximum amount of information around the region of the scale where decisions are being made regarding whether students *below* this score should attend English support programmes. This result is therefore welcomed, as the test does appear to be providing the most reliable information at precisely the point at which it is required.

We may conclude, with some certainty, that for a subtest with only ten items, Section 2 is providing adequate information for the purposes to which the test is being put.

3 Section 3: reading comprehension

Table 5 shows the results of the Rasch analysis for Section 3. Only item 17 was found to misfit, and this item was therefore removed

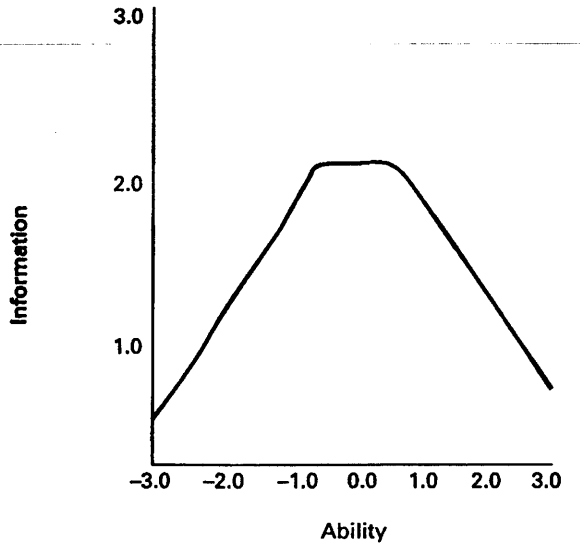

Figure 2 Test information curve for Section 2 (structure)

Table 5 Rasch analysis of Section 3

Item	Difficulty	Standard error	X^2
11	–2.252	0.127	10.190
14	–1.501	0.108	10.459
12	–0.730	0.097	4.660
16	–0.260	0.094	3.983
17	0.016	0.093	14.989 **misfit**
18	1.436	0.106	5.032
13	1.569	0.109	9.180
15	1.723	0.112	8.240

from this form of the test. With the test centred on item difficulty, mean item difficulty was 0.00, and the standard deviation of difficulty 1.48. Average ability was –0.01, with a standard deviation of 1.22. The test characteristic curve for Section 3 is presented in Figure 3. The curve in Figure 3 is not as steep as that in Figure 2, but this is only to be expected, with only eight questions in the subtest. It is difficult to achieve reasonable discrimination with so few items in a subtest. Similarly, the reliability coefficient (equivalent of KR20) is only 0.59. The test information curve in Figure 4 is much flatter than the curve in Figure 2. In an ideal world, the test should be lengthened, as the discussion of this subtest in section VI shows.

The test has been found to be reasonably reliable for its purpose. However, the reliability coefficients do not meet the levels which

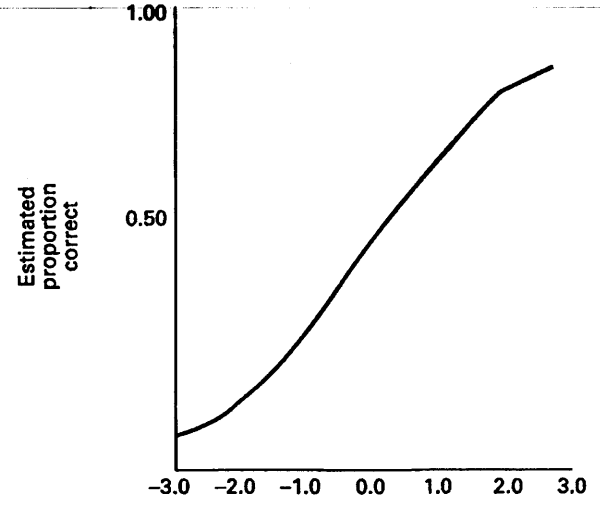

Figure 3 Test characteristic curve for Section 3 (reading)

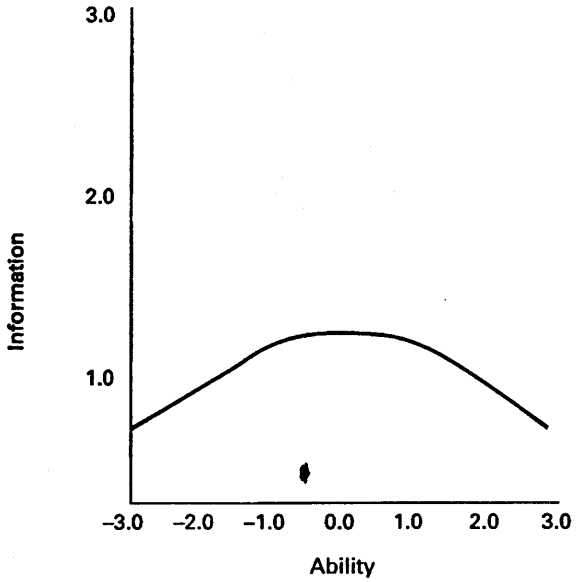

Figure 4 Test characteristic curve for Section 3 (reading)

would be required for a high-stakes test, and this appears to be a function of test length. The relationship between reliability and test length may be described as:

$$n = \frac{rttd(1 - rtt)}{rtt(1 - rttd)}$$

where n is the number of times the test length must be increased with items or raters to achieve the desired reliability, *rttd* is the desired level of reliability and *rtt* is the present observed reliability of the test. For Section 2, to achieve a reliability coefficient of 0.80, it would be necessary to increase test length by 2.34 (28 questions), and for Section 3, 2.77 (22 questions). This would essentially double the administration time of the test!

Compromises need to be made between desired levels of test reliability and the lack of time for testing within a large educational organization. Such decisions cannot be taken by the test designers alone, but need to be discussed at all levels within the institution in relation to general testing policy, including the assessment of the consequences of the unreliability which the institution is prepared to tolerate.

VI Validity

1 Correlational and principal components analysis

The test was designed to measure three separate abilities: writing, English structure and reading. If this is actually the case, the correlation coefficients between the three sections should be modest, and any attempt to factor the matrix should show that subtests load on different factors. Table 6 provides the correlation matrix in the bottom triangle, and the reliability coefficients of the subtests in the diagonal. The upper-right triangle presents the correlation coefficients corrected for attenuation (taking unreliability of measures into account). Although all are significant (to be expected, in subtests which are all language related), variance overlap is not so high that we would wish to challenge the view that each subtest is tapping some unique ability as well as a common ability. It should also be noted that each of the two writing tasks also appears to be tapping different skills to some degree.

The correlation matrix was subjected to principal components analysis (PCA), in order to discover if the tasks were loading on different factors. Eigen values for the extraction of factors were set at 0.6, on the assumption that some of the factors in which we are interested are indeed, as Farhady (1983: 18–19) argues, less than unity. Three factors were extracted, and rotated using the Varimax technique. The results are presented in Table 7. The three-factor solution is clearly interpretable in relation to the correlation matrix. Factor 1 is a writing factor, factor 2 is a reading factor and factor 3 is a structure factor, which the argumentative essay also loads on. However, it must be stressed that this is not sufficient evidence upon which to claim construct validity. Exploratory analysis of this type is suggestive of further avenues for investigating construct validity, but can never be sufficient to meet the requirement for convergent and divergent evidence.

Table 6 Correlation matrix showing association between tasks

	Writing task 1	Writing task 2	Structure	Reading
Writing task 1	**0.87**	0.49	0.36	0.44
Writing task 2	0.43	**0.87**	0.61	0.42
Structure	0.27	0.45	**0.63**	0.44
Reading	0.17	0.30	0.27	**0.59**

Table 7 Varimax-rotated factor solution

	Factor 1	Factor 2	Factor 3
Writing 1	**0.950**	0.049	0.099
Writing 2	**0.552**	0.210	**0.579**
Structure	0.087	0.110	**0.942**
Reading	0.087	**0.983**	0.145
Per cent of total variance explained by rotated components	30.572	25.615	31.363

2 Analysis of cut scores across referred and nonreferred students

Cut scores for referrals were decided on the basis of evidence from the pilot study, and implemented in an operational use of the test. It was therefore imperative that the judgements made be considered in the light of real outcomes. Tables 8 and 9 give the descriptive statistics for nonreferrals and referrals respectively, by task and total test score. The first thing to ensure is that the cut scores established in the pilot study did in fact divide the test-taking population into two

Table 8 Descriptive statistics for nonreferrals by task and total

	Writing 1	Writing 2	Structure-L	Reading-L	Total (raw)
No. cases	1184	1184	1029	1159	1184
Minimum	0	0	−0.64	−2.77	0
Maximum	9	9	2.95	2.74	33
Range	9	9	3.59	5.51	33
Mean	6.92	6.85	1.85	0.63	26.57
SD	2.14	1.01	0.88	1.10	3.25
Standard error	0.06	0.029	0.03	0.03	0.10

Table 9 Descriptive statistics for referrals by task and total

	Writing 1	Writing 2	Structure-L	Reading-L	Total (raw)
No. cases	435	435	408	416	435
Minimum	0	0	−2.67	−2.77	2
Maximum	7	8	2.95	2.74	29
Range	7	8	5.62	5.51	27
Mean	5.31	4.92	0.88	−0.35	20.24
SD	1.24	1.66	1.21	1.13	4.13
Standard error	0.06	0.08	0.06	0.06	0.20

Table 10 Probability of real differences on subtests between referrals and nonreferrals with 1 df (Bartlett test for homogeneity of group variances)

Subtest	Writing 1		Writing 2		Structure-L		Reading-L		Total (raw)	
	X^2	p	X^2	p	X^2	p	X^2	p	X^2	p
Value	154.8	0.000	178.6	0.000	67.38	0.000	0.52	0.471	38.73	0.000

Table 11 Referrals by scores on Writing 1

x	0	1	2	3	4	5	6	7	8	9	O	T
y	10	2	2	13	**36**	**144**	**196**	**32**	0	0	0	435
n	0	0	0	0	**4**	**30**	**241**	**709**	169	21	10	1184
t	10	2	2	13	**40**	**174**	**437**	**741**	169	21	10	1619

statistically distinct groups, and this is shown to be the case in Table 10 for Sections 1 and 2 of the test, and the total test score.

Section 3, the reading subtest, fails to discriminate adequately between referrals and nonreferrals. Although Figure 4 shows that information is being provided at the appropriate point on the scale, this is not enough to make decisions. It is clear that in the case of this subtest, the use of only eight items seriously affects the validity of the subtest for its intended purpose.

Showing that there is a statistical difference between referrals and nonreferrals in Sections 1 and 2 of the test does not, in itself, tell us that the cut score has been adequately placed. In the following four tables (Tables 11–14) are the numbers of students who are referred (y), the number not referred (n) by raw score (x), and the total number of students with each score, for the two writing tasks, structure

Table 12 Referrals by scores on Writing 2

x	0	1	2	3	4	5	6	7	8	9	O	T
y	32	0	6	8	47	**151**	**168**	**22**	1	0	0	435
n	0	0	0	2	1	**23**	**255**	**694**	178	19	12	1184
T	32	0	6	10	48	**174**	**423**	**716**	179	19	12	1619

Table 13 Referrals by scores on Structure

x	0	1	2	3	4	5	6	7	8	9	10	O	T
y	9	2	12	16	29	**39**	**77**	**98**	**80**	**56**	17	0	435
n	0	0	0	0	1	**10**	**74**	**230**	**375**	**338**	146	10	1184
T	9	2	12	16	30	49	151	328	455	394	163	10	1619

Table 14
Referrals by scores on Reading

x	0	1	2	3	4	5	6	7	8	O	T
y	16	**23**	**69**	**109**	**111**	**76**	**25**	4	2	0	435
n	0	**10**	**40**	**151**	**274**	**361**	**230**	92	15	11	1184
T	16	33	109	260	385	437	255	96	17	11	1619

and reading, respectively. Students who did not answer a question are counted as 'omitting' (O). Score ranges where numbers are highlighted in **bold** represent the areas in which there is potential for errors of judgement to be made when making referrals.

The cut score for tasks 1 and 2 (Section 1) was 6, for Section 2, a raw score of 6 (or ability level 0.38) and Section 3, a raw score of 4 (or ability level 0.02). The overall cut score was 22. Raters were asked to consider the evidence from each individual section in making a decision regarding referral. For example, if the overall score was 22+, and on only one of the tasks did the score fall below the cut point, then they were probably not to refer the student. This accounts for some of the nonreferrals in band 5 on both the writing tasks, a score of 5 on Section 2 and a score of 2 or 3 on Section 3, as well as those who were referred within a number of bands above the cut score.

In the case of Sections 2 and 3, however, we are able to calculate the standard error of the cut score, because of the methodology which has been used. The standard error associated with an ability level of 0.02 is 0.852, meaning that we can be 95% confident that a scaled cut score of 0.02 is 0.02 ± 1.67, or anywhere from -1.65 to 1.69. That is a raw score from 3 to 7. The standard error increases as one moves further from the mean! In Section 3, the standard error associated with an ability level of 0.38 is 0.727, which means that we can be 95% confident that a score of 0.38 is 0.38 ± 1.43, somewhere between -1.05 to 1.81, or in raw score terms, anywhere from 3 to 6. However, in practice, we can see that the potential for error is much higher in Section 3 (Table 14) than in any other subtest, as it fails to discriminate between students as a result of its length.

It is only when these calculations are made that the reliability coefficients take on meaning. It highlights the fact that decision-making on cut scores involves error, and that sensitive interpretation of scores on all tasks is required before referring or not referring a student to an English language support programme. However, this evidence lends support to the current practice of interviewing all referred students, to ensure that some students do not needlessly attend the in-sessional programme if it is not necessary. It also highlights the importance of

increasing test length when possible, within the administrative and logistic constraints of large organizations.

One further point does need to be made. There is as yet no way of discovering whether nonreferrals have mistakenly been so categorized, unless they self-refer, or are referred by their subject tutors. These numbers should, however, be low, as the policy of the ELI is: if in doubt, refer. It is easier to correct such an error at a later stage.

3 Concurrent validity

Concurrent validity was investigated by considering the association of scores on the placement test, with scores on the TOEFL. A total of 33 students had taken the TOEFL test immediately prior to arrival at the university, and this proximity allows some degree of comparison between the results of the two tests. Table 15 gives the correlation coefficients between the TOEFL scores of the students, the total score on the placement test and each of the components of the placement test.

From this strength of association between the total placement test score and TOEFL, the best prediction of the placement test score is:

$$\text{Placement test} = -34.73 + 0.1\ (\text{TOEFL})$$

Using the cut score established for the placement test, an estimated TOEFL score of 555 would be required before a student could follow a university course without the need for additional English language support. However, a correlation of 0.64 is clearly not high enough to be able to take these kinds of decisions without the use of the placement test itself. This was confirmed by an examination of the scatter plot of the scores of the 33 students, which revealed two outliers with scores of 510 and 577 on TOEFL, and placement test scores of 12 and 9 respectively. Once these students are removed, the TOEFL–TOTAL correlation is only 0.56, providing the best prediction at:

$$\text{Placement test} = -5.4 + 0.6\ (\text{TOEFL})$$

Table 15 Correlation between TOEFL and the university placement test

Placement test	Section 1 (writing)	Section 2 (writing)	Section 3 (structure)	Section 4 (reading)	Total
TOEFL	0.34 (n/s)	0.57	0.58	0.63	0.64

The new estimate for the lowest TOEFL score required to follow a course of study without English language support would be 499, which appears to be inappropriate on experiential grounds. In conclusion to this section, only moderate association was discovered between the placement test and the TOEFL. This is most likely associated with the small sample size, but may also be a factor of the difference in content and purpose between the two tests. Nevertheless, it is hoped that concurrent validity may be further investigated as additional data are gathered over a number of years, producing a much larger sample of students from whom estimates may be made.

4 Content validity

Content validity was investigated by asking three informants, one each from the Faculties of Human Studies, Engineering and Physics, to 'eyeball' test items. In only one case did an informant feel particularly strongly about a test item, which is given here in full:

Text

Hints of a revolution in superconductivity have been found by French researchers who claim to have discovered a substance that loses all electrical resistance at only 20 degrees Celsius below zero. Superconductors today need to be cooled to −135°C. Superconductors are materials whose resistance to an electrical current disappears completely. Superconducting magnets could make trains levitate and drive ships, and superconducting cables could produce a much more efficient national grid network. Superconductivity was first noted in metals cooled to within a few degrees of absolute zero (−273°C). Then in 1986 ceramics that could superconduct at much higher temperatures caused a storm in the scientific community, holding out the hope that other materials would superconduct at room temperatures. Until now, the best efforts have succeeded only in producing a material that superconducts when cooled with liquid nitrogen – which is commercially useful.

Question

Research into superconductivity is being undertaken because

(a) the use of liquid nitrogen is not practical in non-commercial applications
(b) the railway systems of the world wish to levitate their trains
(c) ceramics is a current topic of discussion within the French scientific community
(d) superconductors only operate at low temperatures

From the pilot study, the item statistics given in Table 16 were obtained, and the item included as a result. However, the expert judge from the Faculty of Physics argued that this item would be unfair to anyone taking the test with scientific training. The reason for undertaking research into superconductivity, he argued, was a combination of (a), (b) and (c). Proposition (d), which he recognized as the required answer, is true, but without the fact that liquid nitrogen is impractical (a), it would not matter that (d) was true. Further, if proposition (b) about the industrial applications were not important, then

Table 16 Item statistics

Item statistics					Alternative statistics					
Item no.	Scale item	Proportion right	Discrimination index	Point biserial	Alternative	Proportion of total	Low	High	Biserial	Key
16	0–6	0.55	0.54	0.52	1	0.20	0.23	0.11	–0.13	
					2	0.10	0.17	0.03	–0.19	
					3	0.08	0.13	0.02	–0.17	
					4	0.55	0.29	0.83	–0.52	*
					O	0.08	0.00	0.00	–0.39	

Note: O = other.

there would be no motivation (other than academic interest, which was not given as an option) to do the research. The main problem with using option (d) as the key, for this informant, was that although the text indicates that it is true, there is no evidence to suggest that it is the prime motive for the research interest. Most scientists are not interested in developing high-temperature superconductors just for the sake of having something that operates at high temperatures, but because they wish to levitate trains, and the like.

Although it is possible for language testers and applied linguists to write good items which are relevant to the disciplines of the students who are taking placement tests, this exercise has provided clear evidence that it is important for subject specialists to be consulted regarding the way in which they would interpret texts as 'insider' readers. Failure to do this may lead to building bias into test items at the construction phase, which may not be detected in the pretesting phase without conducting time-consuming differential item function studies.

5 Feedback from students

After the tests had taken place, a questionnaire was circulated to all students who had been referred for English language support. Of the 435 questionnaires distributed, 71 were completed and returned. This is a response rate of 16.32%, indicating that the following results must be treated with some caution.

The questionnaire requested students to indicate, first, whether they thought the in-sessional placement test was a 'fair' test of their ability to operate in English in an academic context, measured on a five-point Likert scale, and secondly whether they had attended any of the English courses within the in-sessional programme at the English Language Institute. There were open-ended prompts to discover why students thought the test was 'unfair', if they did, how it could be improved, and to find out why referred students had not attended an English course, if they had not. The questionnaire was circulated to all referred students two weeks after the date of the placement test. By this time, referred students were expected to have enrolled in appropriate courses, and so the results of the questionnaire could be compared directly with course attendance.

a Perception of 'fairness': Table 17 indicates that, generally, students did perceive the in-sessional placement test to be fair, with a mean score of 3.1 on the Likert scale. The raw scores are presented in Table 18, to show that of the 71 respondents, one individual did not answer the question, and only 14 thought that the test was either 'very unfair' or 'unfair'.

Table 17 Student perception of test fairness

n	Minimum	Maximum	Mean	SD	Standard error
71	1	5	3.1	0.92	0.11

Table 18 Responses to the question

No response	Very unfair	Unfair	Fair	Quite fair	Very fair	Total
1	4	10	35	17	4	71

Of the 14 respondents who thought that the test was unfair, five provided no reason for their view. The responses for the other nine students were as follows:

Student 1: Greek	More time was needed to complete the test
Student 2: British	The testing environment was poor: the room was too hot and there was no air conditioning
Student 3: Bosnian	The test should have been based on materials which students had already studied
Student 4: British	More time was needed to complete the test, and the questions were 'ambiguous'
Student 5: Japanese	There should be a listening component
Student 6: French	The use of multiple-choice questions should be avoided, as these are not capable of testing proficiency in English
Student 7: Norwegian	The questions were ambiguous. The student also commented that the questionnaire was badly designed, and he could not understand any of the questions
Student 8: Japanese	The invigilator turned up very late on the day of the test, and test administration was poor
Student 9: Singaporean	There should be a vocabulary component

It is clear from these responses that dissatisfaction was, for the most part, not a function of the test itself, but of the restrictions which are imposed on the testing process by the time and facilities available for

testing. The complaint of 'ambiguity' in the questions can be taken as a comment on the distracters in the multiple-choice questions. In a number of the questions the distracters operated extremely well, and some students who got the answers incorrect complained bitterly.

Students who reported that they found the test fair, quite fair or very fair, also echoed some of these views. Thirteen students expressed the need for more time to answer the questions, but at least half these also thought that the test should be longer too. In particular, three students requested additional writing tasks, five students requested a speaking test (which is practically impossible for the entire test-taking population) and four students requested a listening test. One student wished to have a subject-specific test module (business). The only other observation which was echoed, was that of the poor test-taking environment, particularly the lack of ventilation in rooms allocated for the test. There is some concern over the appropriateness of rooms allocated for testing, particularly for those students whose primary language is not English, as it is well known that environmental conditions can have a negative impact on test scores (Ascher, 1990), and those who supply rooms should be made aware of this for future administrations.

b Attendance on the in-sessional programme: The response to the question of whether students had attended English language programmes was a simple yes/no choice. The results are compiled in Table 19 by their response to the first question (Q1 = perception of fairness of the test). A statistical link cannot be tested, because the number of entries in cells is less than the proportion which would generally be considered acceptable to provide reliable p values.

Of the 20 students who were referred to English language support, did not attend an English programme and replied to the questionnaire, 18 provided reasons for nonattendance:

Table 19 The number of students attending English language programmes by response to the question on the perceived fairness of the placement test

Response to Q1	Did not attend	Did attend	Total
No response	1	0	1
Very unfair	3 (75%)	1	4
Unfair	2 (20%)	8	10
Fair	6 (17%)	29	35
Quite fair	6 (35%)	11	17
Very fair	2 (50%)	2	4
Total	20 (28%)	51	71

- Not enough time because of subject-specific study commitments (eight students).
- Claimed they did not know about the in-sessional programme (three students).
- After the oral interview, they were not required to attend (three students).
- Does not know how to use the library, and spends all free time looking for subject-specific materials (one student).
- Attending English classes is a waste of time (three students).

In the last case, there appear to be two different attitudes. In the case of one student whose primary language was English, he was 'offended' at the suggestion that he may need additional support. In the other two cases, the number of years of English as a second language study was seen to 'exempt' them from the need for further language study. One student wrote: 'I don't think because is useful except essay writing. I spent too many years learning English.'

Attendance at in-sessional programme courses is generally very high, and the number of students referring themselves is higher than the number of referrals who do not attend. Total attendances for the in-sessional programme in the 1994/95 academic year reached 511 students. For an optional programme, this seems to be a reasonably satisfactory state of affairs.

VII Discussion

1 Logistic and administrational constraints

Within any large institution there will be logistic and administrational constraints. What is not often recognized, however, is that these constraints lead to limitations on testing, which have a direct impact on reliability of score interpretation. It is therefore important that institutions make policy decisions which take this relationship into account. It should never be assumed that decisions taken by those in administration do not have academic impact. In the case of the University of Surrey, a practical compromise has been reached in the form of a two-tier testing system, with the less precise placement test screening out students from the oral interview stage of the process. This has clearly had a negative impact upon the usefulness of the reading subtest, which needs to be lengthened.

2 The need for pretesting

This study has highlighted the importance of rigorously pretesting all items for inclusion on placement tests. Even with trialling, some

unsuitable items will survive. Without pretesting and *post hoc* analysis, no institution could be sure that its placement tests were providing better information than tossing a coin 1619 times. It is important that institutions such as universities know the estimates of error in their placement procedures, in fairness to the students, and to improve screening and language support provision.

3 Equating test forms

The methodology used in this study will allow forms of the test to be equated, using anchor items. This means that test results will be comparable from year to year, and the English Language Institute will be able to monitor the English language ability, as defined by the test, of incoming students in the future. This will allow it to alert the university authorities to sudden or gradual changes in English language support requirements. It is very common for teachers and students to make comparisons across tests or test formats. In many cases this is not possible, but when forms are equated, it is both possible and beneficial.

Section 1 tasks cannot be equated in the same way as Sections 2 and 3. In the case of Section 1 it will be necessary to link writing prompts using expert judgements, ensuring that prototypical answers awarded at each band of the rating scale are available for rater training. From form to form, much more care will have to be taken with the interpretation of Section 1, even though its reliability in the form discussed here is higher than Sections 2 or 3. This type of equating is essentially judgmental, and one of the less exacting forms of linkage.

In the case of Sections 2 and 3, a strong form of test equating may be attempted, involving strong statistical linking.

4 Future research

In the next academic year, a further form of the test will be produced, and equated with the first form, using a single-group design with concurrent validation (Petersen, Kolen and Hoover, 1989: 256). In principle, this could be done each year until enough forms have been developed to ensure test security even if one form were compromised. As scores on all forms would be strictly comparable (Linn, 1993: 85) it would not matter which form students take. This would add to test security, while maintaining score interpretability. However, this project may be somewhat ambitious. Equating tests is much more difficult with short tests than longer tests, as the burden of information provided by each individual item is much higher in shorter tests (Linn, 1993: 88).

If, however, we are able to produce multiple forms, it also becomes possible to conduct score gain studies. It is a truism that in British universities there is no evidence to suggest that after a certain amount of time on any given programme, a student will 'improve' by '*x*', whatever the unit '*x*' is in terms of ability. Indeed, score gain studies, even if conducted, would be meaningless, without a clear understanding of what '*x*' is. In this case, the unit of measurement is the scale which has been established in the October 1994 study. This is arbitrary, but person and item free (Masters, 1990). As such, score gain studies may be conducted with new groups of students over different timescales, following different courses. If successful, this would allow the English Language Institute to say to a department that a particular student would require (given error) from '*a*' to '*b*' months of language tuition to reach a level at which he or she would, with *p* probability, be able to cope with an academic course with (or without) English language support.

This kind of information is not currently available, but in principle could be, as the result of a careful development of research within the context of specific language programmes of large educational institutions.

VIII Conclusion

We are aware that there are other approaches to developing placement tests, many of which are effective where the numbers of students taking the test are small and the constraints in local administration allow the use of lengthy instruments which include oral components (Paltridge, 1992). Each approach must be sensitive to the constraints imposed by, and the information requirements of, unique institutions. Similarly, when conducting score gain studies in the future, criterion-based approaches such as that suggested by Brown (1989) will be considered, especially if multiple forms of the current test are available. We nevertheless believe that the current test fulfils its purpose as well as can be expected within the context described in this article.

The methodology used in developing the evaluation of this placement test was based on Wall, Clapham and Alderson (1994), as the first article in the field of language testing to address the issue of the assessment of placement instruments in any depth. However, the methodology used in this study includes additional aspects which would appear to be worthy of investigation in the particular context of placement testing. It is hoped that the approach adopted, building on Wall, Clapham and Alderson's pioneering work, may be of use to others in the evaluation of their placement tests.

IX References

APA 1985: *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ascher, C. 1990: *Assessing bilingual students for placement and instruction. ERIC Digest* 65, ED322273. New York: ERIC Clearinghouse on Urban Education.

Assessment Systems Corporation 1994: *RASCAL 3.5. Rasch analysis program*. Minnesota, MN: ASC.

Brown, J.D. 1989: Improving ESL placement tests using two perspectives. *TESOL Quarterly* 23, 65–83.

Crocker, L. and **Algina, J.** 1986: *Introduction to classical and modern test theory*. Chicago, IL: Holt, Rinehart & Winston.

Farhady, H. 1983: On the plausibility of the unitary language proficiency factor. In Oller, J.W., editor, *Issues in language testing*, Rowley, MA: Newbury House, 11–28.

Goodbody, M.W. 1993: Letting the students choose: a placement procedure for a pre-sessional course. In Blue, G.M., editor, *Language, learning and success: studying through English*, London: Modern English Publications and the British Council, 49–57.

Linn, R.L. 1993: Linking results of distinct assessments. *Applied Measurement in Education* 6, 83–102.

Masters, G.N. 1990: Psychometric aspects of individual measurement. In de Jong, H.A.L. and Stevenson, D.K., editors, *Individualizing the assessment of language abilities*, Clevedon: Multilingual Matters, 56–70.

Paltridge, B. 1992: EAP placement testing: an integrated approach. *English for Specific Purposes* 11, 243–68.

Petersen, N.S., Kolen, M.J. and **Hoover, H.D.** 1989: Scaling, norming and equating. In Linn, R.L., editor, *Educational measurement*, New York: National Council on Measurement in Education and Macmillan, 221–62.

Popham, W.J. 1978: *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Schmitz, C.C. and **DelMas, R.C.** 1990: Determining the validity of placement exams for developmental college curricula. *Applied Measurement in Education* 4, 37–52.

Shohamy, E. 1983: The stability of oral proficiency assessment in the oral interview procedure. *Language Learning* 33, 527–40.

—— 1988: A proposed framework for testing the oral language of second foreign language learners. *Studies in Second Language Acquisition* 10, 165–79.

—— 1990: Language testing priorities: a different perspective. *Foreign Language Annals* 23, 385–94.

Shohamy, E., Reves, T. and **Bejarano, Y.** 1986: Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal* 40, 212–20.

Truman, W.L. 1992: College placement testing. *AMATYC Review* 13, 58–64.

Upshur, J.A. 1971: Objective evaluation of oral proficiency in the ESOL classroom. *TESOL Quarterly* 5, 47–59.

van Weeren, J. 1981: Testing oral proficiency in everyday situations. In Klein-Braley, C. and Stevenson, D.K., editors, *Practice and problems in language testing. Volume 1*, Frankfurt: Bern, 54–59.

Wall, D., Clapham, C. and **Alderson, J.C.** 1994: Evaluating a placement test. *Language Testing* 11, 321–344.