

## **THE PREDICTIVE VALIDITY OF A READING TEST FOR ENTRY TO MAINSTREAM COURSES**

Jeannette Watts

Languages Department

School of Communication and Tourism

Wellington Polytechnic

### **Introduction**

In her investigation into the predictive validity of the IELTS test Bellingham demonstrated that diagnosis, prior to entry, of a potential student's proficiency in English provides valuable information from which the student and academic advisers can negotiate a pathway to academic success (1995, p.27). This paper reports on a similar undertaking to take the guesswork out of negotiating such a pathway. It traces the test development process of writing, piloting, analysing responses, trialing, testing for reliability and validity, to the final stage of testing its usefulness as a predictive tool.

The need for such a test arose out of the desirability to provide ESOL students with an indication of the likelihood of managing the reading material of their mainstream courses at Wellington Polytechnic. Recently arrived immigrants with a range of work experience and English language skills, together with those who have spent some years in the New Zealand school system, apply for courses, particularly business related courses, often with inadequate indication of their English language ability. The purpose of this test therefore is to enable applicants to make more informed choices upon enrolment.

### **Test Development**

Several issues needed to be addressed throughout the process of test development:

#### **One test or subject specific tests?**

The first question to be considered was if the test should be subject specific. According to Clapham (1993) students did not score significantly higher on reading tests within their own subject area and so were not disadvantaged by taking a test module outside their area. She also suggests that higher level students may be more successful at inferring meaning from the context while elementary learners rely more on background knowledge.

Further to this, if the time consuming task of developing separate subject specific tests were pursued, it would be necessary to carry out the difficult undertaking of ensuring an equal degree of difficulty across different versions of the test.

### Text effect

Related to the issue of subject specificity is the choice of text used in the test and its degree of familiarity to test takers. In selecting a text, a balance needed to be achieved between its being neutral in terms of subject bias (as far as that is possible) and its being reasonably familiar to all, thus limiting the effect of the text on test performance.

The issue of text effect has not been dealt with in depth in the literature in the past, one reason being that there has been no clear consensus on the role of texts (Allerson and Grabe, 1986, p.164). They report that there are three major concerns: whether texts should be on widely familiar or unfamiliar topics, whether more but shorter texts are better than fewer but longer texts, and whether texts should be of a general topical nature or more related to specific subject areas. The objectives of programmes, they conclude, should determine the kinds of texts used. In regard to text length, for example, if students are to identify main points from subsidiary detail, a text of at least one page would be necessary to enable the testing of this skill.

Alderson highlights the dilemma in regard to topic. In view of the importance of a reader's schemata in understanding the content of the text, a test should acknowledge and reward the application of it to a text. Yet the purpose of a reading test is to assess reading skill, not knowledge of the topic. To avoid the risk of bias towards some test takers, Alderson suggests choosing a number of texts on a range of different topics and a variety of different types of text (1996, p.221). While this is clearly a valid point, the implementational constraints of this test prevent it. It is foreseen that only a limited time will be available for administration of the test to individuals, so it would be feasible to have two or three different texts only if they were very short. However, the objectives of the mainstream courses indicate the importance, for example, of identifying main points from subsidiary detail, which requires a longer text. So just one page-long text was chosen, supplemented with a related table.

In view of the research on the effect of prior knowledge and reading tests, it was desirable to select texts dealing with topics equally familiar, or equally unfamiliar, to all test takers. As the latter was likely to lead to a choice of an obscure topic, the former was chosen.

Largely for security reasons the decision was made to develop parallel forms of the test (Version A and Version B), making it therefore necessary to have two parallel texts. Eventually two articles from the *New Zealand Business Journal* were selected. There were several reasons for the choice:

- Each article was a complete text and of appropriate length.
- Although they were from a business journal the articles were general in nature.
- The topics both related to migrants in New Zealand so presumably would be of some interest to test takers.

- The articles contain a mixture of formal and less formal language, which is reasonably typical of the case study approach which many courses pursue.
- Both articles were of similar style and text type, and were found to have similar readability statistics and levels of vocabulary.

The topic in both articles is concerned with Asians in New Zealand. This may be seen to present a bias in favour of Asian test takers, but given that there has been considerable media coverage of Asian immigration to New Zealand, this seems to be a reasonable decision.

### **Test content**

In order to ensure the test contained the range of reading skills considered necessary for potential course members it was essential to seek the input of subject lecturers by means of a 'de-jargonised' questionnaire (Appendix 1). To develop this questionnaire, an examination was carried out of the learning outcomes and New Zealand Qualifications Authority unit standards of seven major Wellington Polytechnic programmes, to identify the reading skills required in these courses. By considering these reading tasks in conjunction with the reading sub-skills listed by Nuttall (1996, pp.69-125) the questionnaire was developed and sent to ten lecturers whose courses, taken from the seven programmes initially examined, typically had a significant proportion of ESOL students. They were asked to assign relative weighting to reading subskills, the results of which were subsequently averaged out to present an overall picture of priorities.

The decision about which skills to include in the test and the proportion of items, or weighting, was based on these reported priorities. Weighting refers to the extra value to items because of their greater importance. This however rarely leads to increased reliability or validity. Alderson et al (1995, p.52) quote Ebel (1979) in condemning differential weighting:

If an achievement test covers two areas, one of which is judged to be twice as important as the other, then twice as many items should be written in relation to the more important area. This will result in more reliable and valid measures than if an equal number of items is written for each area and those for the more important area are double-weighted.

In view of this, each item of this test is worth one mark.

### **Test format**

There were two important concerns to be addressed in regard to test format. The first of these was 'face validity': that is, the items had to be familiar to students if they were to believe it was a fair test and accept its outcome. Secondly, because the test was designed to be administered by subject lecturers with little or no chance for inter-rater checks, the items had to be objectively marked items.

This practical concern of no inter-rater checks and the subsequent need for objectively marked items had to be seen in light of the tension between the demands of reliability and validity. While objectively marked tasks may be perceived to produce consistent measures, they might not represent valid reading tasks. They also may not represent reading ability, and produce higher results, thereby actually being less reliable. Clearly the method used may affect the student's score, so the influence of task or method effect should be reduced as much as possible.

But what is known about format effect on student performance? Although research shows some evidence of format effect, with researchers asserting that some tasks by their nature influence the interaction between the reader and the text as well as the reader's performance on a reading comprehension test, Weir writes:

Given the limited state of knowledge concerning the effect of test formats, the only practical approach at present is to safeguard against possible format effect by spreading the base of a test more widely through employing a variety of valid, practical and reliable formats for testing each skill.

(1990, p.45)

This is supported by Alderson et al (op cit, p.45) who describe knowledge about test method effect as 'rudimentary'.

My test aimed to reduce format effect by utilising a range of tasks including: sentence completion, evaluating the truth value of statements, identifying referents, selecting multi-choice vocabulary synonyms, a selective deletion cloze to summarise the article, chart completion to illustrate the relationship between main and subsidiary ideas, and a yes/no task to demonstrate recognition of implications.

## **Analysis of test results**

### **Item analysis**

In order to judge how effectively the items in both Version A and Version B functioned, they were trialed and the responses analysed. Version A was piloted on a group of nine learners in an advanced class, and revealed a lack of clarity in the instructions. Both versions underwent some re-writing, after which Version B was piloted on twenty test-takers of similar proficiency level and ethnic mix to eventual test takers. Alderson et al write that for pre-testing a sample of twenty is "a good number" (ibid, p.75).

The results of Version B's pilot were analysed using classical item analysis. This was used because of the convenience of calculating results manually. The alternative, the Rasch model,



would have produced results with an unacceptably high margin of error because the sample was less than one hundred (Alderson et al, *ibid*, p.91).

Two statistics were particularly useful for this analysis, the *item facility value* and the *item discrimination value*. *Item facility* measures the difficulty of each item and is expressed as a proportion of the number of the test-takers who answered it correctly. Its use is in indicating very difficult or very easy items, neither of which are informative of the varying abilities of the group. Items that produce an item facility value of between 0.4 and 0.8 are considered to be the more effective items (Read, 1995, p.20).

To add to the picture of well functioning items, *item discrimination values* were also sought. This statistic indicates how well an item distinguishes between students at different levels of ability. It also shows if both the test and the item are sorting out students consistently and therefore reliably. Because the purpose of this reading test is to separate or discriminate between test-takers, items which discriminate well are obviously desired. Although there are various ways of calculating item discrimination the following formula was used because of its ease of use. (Alderson et al, *op cit*, p.81). It presents a clear distinction between upper and lower groups of students while excluding the average scores in the middle third, and as such is a summary statistic which measures spread.

no. of students in top third who answered item correctly	<i>minus</i>	no. of students in bottom third who answered item correctly
<hr/>		
no. of students in each third		

As a result of this analysis, eight items were eliminated, leaving thirty-five. The major reason for discarding items was that they were either too hard or too easy for most, or did not discriminate sufficiently well. These thirty-five items formed the basis of the second draft of Version B. A second draft of Version A was also written to parallel Version B. The degree of equivalence between the two was subsequently measured with a different group of seventeen students.

### **Descriptive statistics**

These measures allow the comparison of difficulty level and spread of scores of different tests with each other. The three measures of central tendency (the mean, the mode and the median) show how the scores cluster together, while the measures of dispersion (the range and the

standard deviation) show how widely the scores are spread out. Together they show how appropriate the test is for its intended purpose, such as if it is a suitable level of difficulty or if it is capable of discriminating between students. A test such as this reading proficiency test has to distinguish between levels of students, so a very difficult or easy test with a skewed distribution will not be suitable because too many people will be clustered at either extreme. Instead there should be a spread of scores with only a few students getting one particular score. The descriptive statistics are presented in Table 1.

	<u>Test A</u>	<u>Test B</u>
Number of test takers	17	17
Possible score	35	35
Mean	23(65.7%)	21(60%)
Median	24	23
Standard Deviation	7.4	6.2
Variance	54.7	38.44
Range	28	22

**Table 1 : Descriptive Statistics of Tests A and B**

Because the means are not high, it indicates this group of students did not find this test easy.

Before the poorly functioning items were eliminated, the mean had been lower at 64.65% for A and 66.74% for B, indicating slightly easier versions of the test. The gap between the two means widened on the revised version, indicating that the level of difficulty of B increased more than did the level of difficulty in A. That is, test B which was already the harder, became harder still with the removal of the easy or poorly discriminating items. Because the remaining items produced means of 65.7% and 60% for Versions A and B respectively, statistical guidance was sought regarding their equivalence, with the advice that scores would need to be adjusted before considering them truly equivalent tests, but that for practical purposes they were sufficiently close to be considered parallel.

### **Reliability estimates**

Reliability is a primary quality to be considered in developing and using tests. It is concerned with the consistency of measures across different times, test formats, raters, and other characteristics of test measurement. Thus it is a quality of test scores and as such can be affected by factors other than the trait being measured. Reliability therefore can never be perfect but at best controlled for by limiting the effects of sources of measurement error. Then systematic variations are more likely to be measured; that is, changes in ability not changes in external factors are measured.

In this reading test attempts were made to reduce the threats to reliability by:

- ensuring similar administrative procedures will followed
- scoring objectively
- using familiar formats
- trialing for the ability to discriminate
- trialing for the appropriate level
- making the test length as long as practicality allows
- checking of clarity of instructions by several people

Further to this, to minimise the possibility of an ordering effect, a counter balanced design was used when measuring the degree of equivalence between the two versions, thus:

	Day 1	Day 2
Students 1-8	Version A	Version B
Students 9-17	Version B	Version A

Despite these efforts, it can still never be assumed the test score is an exact measure of proficiency, but rather an estimate. The following reliability measures, like all reliability measures, are based on the assumption that has to be made that the sample of seventeen test takers is representative of a random sample of future test takers. These seventeen test takers were representative in as far as they were enrolled in a Study Skills class in preparation for undertaking mainstream study in the very near future.

Three different statistical methods were used to estimate the reliability of the test results: internal or inter-item consistency, the split half method, and parallel form reliability, which is concerned with inconsistencies across forms of the test.

*Internal consistency* was measured in two ways, firstly by the KR-21 formula (Harrison, 1983, p.126):

$$R = 1 - \frac{M(n-M)}{ns^2}$$

where M is the mean, n is the number of items in the test, and s is the standard deviation, as listed in Table 1. The resulting correlation coefficients were:

0.856 (Version A)

0.782 (Version B)

Correlation coefficients are produced to support the evidence of reliability. Correlation can be used as a measure of the consistency between performance on two different tasks. As correlations are defined, they always lie between -1 and +1, and are reached by obtaining two different measures of the performance of the same group of learners and then comparing them.

A correlation figure indicates the extent to which the two measures rank the learners in the same way. The strength of this relationship can be interpreted in relation to the level of statistical significance, which gives an indication of the probability that the correlation coefficient is the result of chance factors. The most widely used level of probability in language testing is  $p < .05$ . (That is, there are less than five chances in a hundred that the correlation is affected by random variation.)

Another procedure to obtain a correlation coefficient is the *split half method*, also referred to as the Spearman Formula (Richards, 1992, p348):  $R = \frac{1-6\sum d^2}{N(N^2-1)}$

(where  $d$  is the difference in rank between scores on odd and even numbered items and  $N$  is the number of test takers). This formula was used because of the relative homogeneity of the skills tested. Alderson et al (op cit, p.89) mention that if the test contains different sections testing different skills, the sections will not correlate highly with each other and the reliability will be lower. The same authors believe that this reliability index produced by the split half method generally produces similar results to the more complex KR20 formula. This formula could not be used confidently because the students were not given unlimited time to complete the test, thereby resulting in some unfinished items by the less able students, which tends to produce a reliability index which is too high (Alderson et al, ibid). The split half procedure was carried out on the results of Version A, producing a correlation between scores on odd numbered items and even numbered items of 0.928.

The third measure used to estimate reliability of test results was *parallel form reliability* which was used because of the necessity to correlate the scores of Versions A and B. Clearly they need to have a high correlation if they are to be used as equivalent tests. Although there are many ways of calculating correlation coefficients, rank order correlation was used because there is only a small number of scores and they are easily ranked. Again the Spearman Formula was used, and produced a correlation coefficient between Versions A and B of 0.734.

Because the number of the sample is small ( $N = 17$ ), quite a large co-efficient is required in order to be sure of the strength of the relationship between the two measures. At 0.734 the correlation can be at moderate to high. It is well above the critical value of 0.488 at  $p < 0.5$ , so the correlation is a statistically significant one.

The overlap between the two measures is expressed by the co-efficient of determination, which is obtained by squaring the co-efficient (Richards, 1992, p59), that is:  $0.734^2 = 0.559$ . Hence there is an almost 56 percent overlap between the two sets of rankings.

While determining the correlation figures between the two tests it became clear a practice effect had occurred. In terms of scores, nine students did better on the second day, while five

did better on the first day and three stayed the same, irrespective of which form of the test they did. The order of the test seemed to have more impact than which version it was: seven students did better on Test A, seven students did better on version B, and three stayed the same.

### **Validity**

Once it had been established that the tests were able to produce consistent, reliable results it was necessary to obtain evidence for the validity of the test. While reliability is a quality of test scores themselves, validity is a quality of test interpretation and use. According to Bachman validity is the extent to which the inferences or decisions we make on the basis of the test scores are *meaningful, appropriate and useful* (1990, p.25). That is, a test result should be meaningful in relation to its intended purpose. The purpose of this reading test is to provide information about the likelihood of success in handling reading material in mainstream courses. Thus it was necessary to determine how to interpret the results. What did it mean to score 20? What did it mean to score 30?

If a test result is to be used as an indication of a particular ability it must measure that ability and nothing else. Therefore a test score that is not reliable cannot be valid: it will not be a meaningful indicator of the ability in question. To put it another way, a reliability estimate examines the variance in test scores themselves while validation examines other sources of variance. So the process of validation extends beyond the reliability to the relationship between test performance and factors outside the test itself.

Just as there is not a clear cut distinction between reliability and validity, nor is there a clear cut valid/not valid distinction. Evidence needs to be gathered then judgements made about the relative validity of a test and the degree of confidence one can hold in it. Nevertheless, data can be gathered to show a test has a degree of validity for the purpose for which it is intended. Evidence can be gathered either internally or externally. Internal validity relates to studies of the perceived content and its effect, while external validity relates to studies comparing students' test scores with measures gained from outside the test.

*Internal validity* includes face validity, content validity and response validity. This study employs the first two of these.

*Face validity* refers to "surface credibility or public acceptability" (Alderson et al, op cit, p.172). In the case of this reading test, face validity is important. If test takers do not believe it to be an authentic test of their reading ability they will be reluctant to accept its results. Some degree of evidence for face validity was gathered by asking all test takers upon



completion of the test for their opinion of it. Most gave extensive responses which overall were positive.

*Content validity* is concerned with the representativeness of the content. It is thus concerned with what goes into the test, its coverage and relevance, and the balance between the items. This is typically achieved by seeking the judgement of experts in the field, although as Alderson and Lukmani's research (1989) shows, experts do not always agree.

For this reading proficiency test, content validity was sought by making use of subject teachers' input about the relative importance of reading skills, and by asking six experienced ESOL colleagues to perform and comment on a formal basis on both versions of the test. Substantial modifications were made on the basis of their feedback.

*External validity* includes concurrent validity, predictive validity and construct validity.

*Concurrent validity*, as the name suggests, involves obtaining two independent measures of ability at about the same time and comparing or correlating them.

For this test, this was done in two ways: firstly by correlating students' ranking on the reading test with teacher ranking of students (made after four weeks of full-time teaching with the class). Using rank order or Spearman's correlation produced the following outcome:

	Version A	Version B
Correlation co-efficient	0.872	0.854
Co-efficient of determination	0.760	0.720

To gather further evidence of validity, Spearman's Formula was used to produce rank order correlations between performance on the reading test and the end-of-course rational deletion cloze test:

	Version A	Version B
Correlation co-efficient	0.735	0.738
Co-efficient of determination	0.540	0.545

According to Alderson et al (op cit, p.178) most concurrent validity correlation coefficients range from +0.5 to +0.7, with higher coefficients being unlikely for measures like teacher assessments. These figures therefore are acceptable.

A second process of external validation was attempted by comparing reading test results with those of the IELTS reading sub-test but the sample size was too small to make any firm conclusions.

The purpose for which this reading test is intended is as a predictor of the likelihood of managing reading material in mainstream Polytechnic courses. The final and most important phase of its development therefore was to test its predictive validity.

*Predictive validity* makes use of a future measure of student performance for validating the test. Hence a proficiency test might be validated by future performance in an academic course, or another test of the ability which the initial test was intended to predict. A high correlation between the two would indicate a high degree of predictive validation.

There are problems with predictive validation, as described by Cattell in 1964 :

The correlation of a test now with a criterion next year has a host of determinants among which the properties of the test may well be insignificant. Future prediction after all requires knowledge of the natural history of the trait, the laws of psychology, and (not the least!) the changing life situations, for example the stock exchange, which will affect the individual in the interim.

(cited in Bachman, op cit, p.252)

With so many intervening factors, it is not surprising therefore that "the predictive strength of an English proficiency test is commonly low, with a correlation of 0.3" (Davies, 1988, p.33).

### **Testing for predictive validity**

A total of twenty-eight participants were drawn from two consecutive language support classes of the Introductory Certificate in Business Skills course, for which International students require an IELTS score of 5.0. A further twenty-two participants were drawn from the English Study Skills class of the New Zealand Diploma of Business (formerly National Certificate in Business), for which International students require an IELTS score of 5.5. Soon after the course began one participant withdrew for health reasons, so it was decided to exclude her from the study, leaving twenty-one participants.

All participants undertook the reading proficiency test in the first week of their respective courses. The results of this enabled two aspects of the research:

- A correlation by means of a rank order correlation between reading test results and final results of their courses (which were both 18 weeks in length.)
- A comparison between results on the reading test and overall pass rates.

### **Results of the predictive validity study**

#### **Group 1 : Introductory Certificate in Business Skills**

Correlation between performance on reading test and final results:

For this group, correlations were made with the combined final results of six core courses, which were chosen because it is essential that students pass these courses. A rank order correlation was done to measure the strength of the relationship between test results and course results. This means of correlation was used because it enables comparison between two different scales. To obtain the correlation coefficient the Spearman Formula was again used:

$$R = 1 - \frac{6 \times \sum d^2}{N(N^2 - 1)}$$

The resulting correlation co-efficient of 0.724 is moderate to high, suggesting that performance on the reading test definitely bears a relationship to performance in the six courses.

Comparison between reading test results and pass rates:

As shown in Table 2, most of those who scored more than sixteen were successful on their course, while most of those who scored sixteen or less did not succeed: that is, they either failed or dropped out.

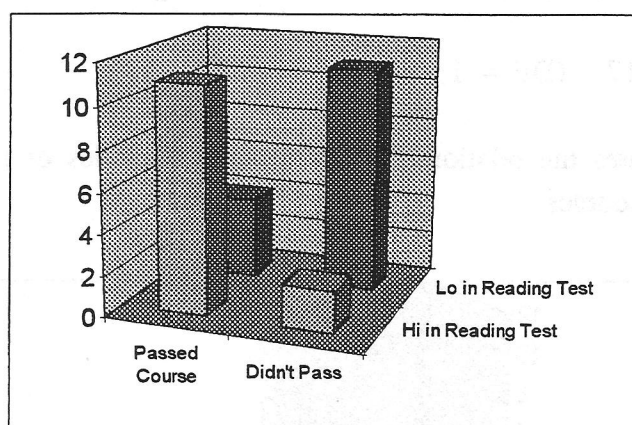
Score on reading test	No. passing course	No. not passing	Total
> 16	11	2	13
≤ 16	4	11	15

**Table 2 : Reading test results and pass rates for Introductory Certificate in Business Skills course.**

A Chi-Square test was carried out to show if the effects were dependent or not, that is, if passing the course related to attaining a reading score of over sixteen, with the following result:

$$\chi^2 = 9.403 \quad (\text{DF} = 1, \text{P-Value} = 0.002)$$

This relationship is demonstrated graphically in Figure 1.



**Figure 1: Relationship between reading test results and success on Introductory Certificate in Business Skills course**

### Group 2 : Diploma of Business

Correlation between performance on reading test and final results:

For this group, correlations were made with the combined final results of the three courses undertaken concurrently: English Study Skills, Computer Concepts, and Business Communication. The correlation coefficient for this was 0.60.

Because the number of this sample is small ( $N = 21$ ) a quite large coefficient is required to be sure of the strength of the relationship between the two measures. For a test population of twenty-one the critical value or the level of statistical significance is 0.5324, so this measure exceeds this.

Comparison between reading test results and pass rates:

Table 3 below shows that almost all of the participants who attained a score of more than twenty passed their three courses. Conversely most of those who scored twenty or less did not pass their three courses.

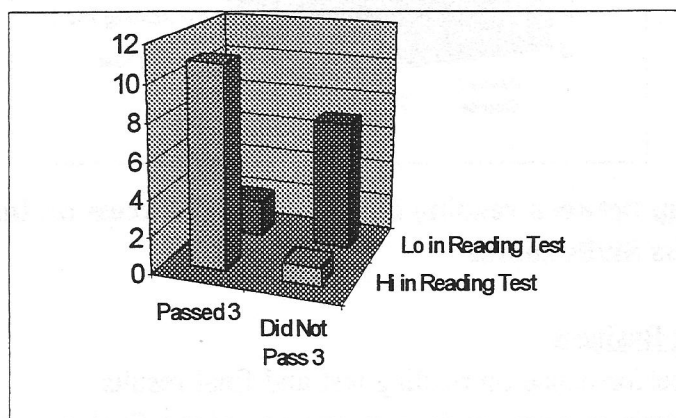
Score on test	No. passing 3 courses	No. passing 2 courses	No. passing 1 course	No. passing 0 courses	Total number
> 20	11	0	1	0	12
≤ 20	2	2	1	4	9

**Table 3 : Reading test results and pass rates for Diploma of Business course.**

A Chi-Square test was performed on the number of those passing three courses and the total number of those not passing three courses, with the following result:

$$\chi^2 = 10.517, \quad (DF = 1, \quad P\text{-Value} = 0.001)$$

Again, the graph illustrates the relationship between the variables of reading test result and success in passing three courses.



**Figure 2 : Relationship between reading results and success on Diploma of Business course**

## Conclusion

The correlations produced between the reading test and course results are high compared to the commonly produced correlation of 0.3. (Davis, op cit, p.33) This may in part be explained by the homogeneity of the test population. This is not to say they were homogeneous in terms of language proficiency, which would lower the correlation, but homogeneous in educational experience in as far as they were all involved in the same course of study, which might have reduced the range of variables influencing their academic performance (Read, personal communication, September 1998). The strength of the correlation indicates a moderately high predictive validity of the test. In other words, it is possible to have confidence in the relationship between performance on the reading test and future performance on the two business courses under study.

In the light of these results, suggested cut-off points for entry into various levels of courses are set at:



0 - 16 = not yet ready for mainstream courses

17 - 20 = ready for certificate level courses

21 - 30 = ready for diploma level courses

31+ = ready for degree level courses

The relationship between the score of 31+ and success on a degree level course has not yet been tested and is at this stage based on observation of students whose reading score was in the 28-29 range. Further research is required to collect stronger evidence of this measure.

This predictive validity study revealed that even on a small sample it is obvious there is a strong relationship between test scores and pass rates. This together with the reliability estimates enables a reasonable degree of confidence in interpreting results of applicants to courses. To increase the sample size, the results of successive intakes of ESOL students into mainstream courses will be added, continuing to build on this initial research.

The purpose of this study has been to develop a test that will provide applicants and lecturers with information about ability to cope with reading material in a course of study. But, as the end-users of the test must always keep in mind, language proficiency is just one of many factors contributing to academic success, particularly at higher levels of proficiency. The outcome of this reading test must be seen as forming just one part of the complex picture.

### **Acknowledgements**

I would like to thank John Read, Victoria University of Wellington and Mike Camden, Wellington Polytechnic for their help.

### **References**

- Alderson, J. (1996). The testing of reading. In C. Nuttall (Ed.), *Teaching reading skills in a foreign language*. (pp.212-228). Oxford: Heinemann.
- Alderson, J. and Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5, 253-270.
- Alderson, J., Clapham, C. and Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

- Allerson, S. and Grabe, W. (1986). Reading assessment. In F. Dubin, D. Eskey, and W. Grabe (Eds.), *Teaching second language reading for academic purposes*. Massachusetts: Addison-Wesley.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bellingham, L. (1995). Navigating choppy seas: IELTS as a support for success in higher education. *The TESOLANZ Journal*, 3, 21-28.
- Clapham, C. (1993). Is ESP testing justified? In D. Douglas, and S. Chapelle (Eds.), *A new decade of language testing research*. Alexandria, VA: TESOL.
- Davies, A. (1988). Procedures in language testing. In A. Hughes (Ed.), *Testing English for university study* (pp.29-35). ELT Doc 127, Modern English Publications.
- Harrison, A. (1983). *A language testing handbook*. London: Macmillan.
- Richards, J., Platt, J. and Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics*. Harlow: Longman.
- Weir, C. (1990). *Communicative language testing*. Hempel Hempstead: Prentice-Hall.

## APPENDIX 1: Ratings by Subject Teachers - Questionnaire

### READING COMPREHENSION SKILLS

*Please rate the importance of the following reading comprehension skills in terms of the reading required by your course. Give each skill a percentage relative to the others, totalling 100%. Some skills may have equal ratings. If there are others you wish to add, please do so.*

#### Meaning of Words

##### 1. Word recognition skills :

- Recognising words similar to those already known
- Guessing the meaning of unknown words from the context
- Locating a specific item of information quickly
- Reading charts, tables, graphs, maps, etc

##### 2. Word attack skills

- Determining word meaning by recognising
  - if the unknown word is a noun, verb, etc
  - prefixes and suffixes e.g. *un-*, *-ness*
  - compound words e.g. *second-hand*

#### Reading for Meaning

##### 3. Understanding long sentences by recognising the grammar and word order of the sentence.

##### 4. Recognising and interpreting words and phrases that link the text together. That is:

- a) Proforms e.g. *it, our, this, those, such, other, same*
- b) Ellipsis (a word omitted) e.g. *The days are hot and the nights cool.*
- c) Lexical cohesion (an idea repeated with a different word) e.g. *The car sped around the corner. It was a beautiful, red vehicle.*

##### 5. Interpreting words and phrases that show relationships between different parts of the text.

- For example:
- *first, next, then, the following day*
  - *in conclusion, for example, to sum up*
  - *incidentally, certainly, more importantly*

#### Reading Between the Lines

##### 6. Recognising the function or purpose of a sentence when the writer has not explicitly stated it, such as if the sentence is a definition, an example, a hypothesis, a contrast, or an explanation.

##### 7. Identifying the way in which a text is organised, such as by chronological order, thus being able to locate specific information more easily.

##### 8. Recognising the writer's assumptions, such as opinions and attitudes, which underlie the text and which the reader is expected to share.

##### 9. Recognising implications and making inferences from the text by drawing unstated conclusions.

##### 10. Predicting outcomes by tracing the development of ideas throughout the text.

##### 11. Evaluating the text by distinguishing between:

- important points and supporting details
- relevant and irrelevant information
- fact and opinion

ARTICLE 1. PURPOSE AND SCOPE

The purpose of this Association is to advance the science and practice of medicine, to promote the highest standards of medical education, and to protect the public interest in the medical profession.

ARTICLE 2. MEMBERSHIP

1. There shall be two classes of members: (a) Regular members, who shall be physicians and surgeons, and (b) Associate members, who shall be persons interested in the advancement of the medical profession.

ARTICLE 3. OFFICERS

The Association shall elect annually a President, a Vice-President, a Secretary, and a Treasurer, who shall hold office for one year and shall be eligible for re-election.

ARTICLE 4. MEETINGS

The Association shall hold an annual meeting, and may also hold special meetings as may be deemed necessary.

ARTICLE 5. FINANCIAL AFFAIRS

The Association shall have the right to acquire, hold, and dispose of real and personal property, and to enter into contracts for the same.

ARTICLE 6. RELATIONS WITH OTHER ORGANIZATIONS

The Association shall be authorized to enter into such relations with other organizations as may be deemed to be in the best interests of the medical profession.

ARTICLE 7. AMENDMENTS

The Association may amend its constitution and bylaws by a two-thirds vote of the members present at a meeting called for that purpose.

The Association shall have the right to sue and be sued, and to defend itself in any court of law or equity.

The Association shall have the right to make and alter its rules and regulations, and to enforce them.

The Association shall have the right to make and alter its financial statements, and to audit them.

The Association shall have the right to make and alter its membership rules, and to enforce them.

The Association shall have the right to make and alter its relations with other organizations, and to enforce them.

IN WITNESS WHEREOF