Do English and ESL teachers rate essays differently?

Kieran O'Loughlin

1. Introduction

An important issue in the direct assessment of writing is that of rater background and experience and, in particular, how this affects both global judgements of writing quality and the perspectives from which raters arrive at these judgements.

This study aims to investigate whether secondary teachers of English as a mother tongue and of English as a second language (hereafter referred to as "English" and "ESL") rate essays differently. Using recently collected data it examines the reliability between and within these two groups of raters in assessing the same set of forty essays written by both English speaking and non-English speaking background (also abbreviated as "English" and "ESL" respectively) final year secondary students. As the method of scoring used was both holistic and analytical, it also attempts to investigate which factors such as content, organisation or syntax, most strongly influence the global assessment of these essays.

2. Background to the study

There has been a surprisingly limited amount of research comparing the direct assessment of writing by English and ESL teachers given the amount of common ground they share as teachers of writing.

Morgan (1990) compared the global scores assigned by a group of seven English and ESL teachers of sixteen essays written by final year secondary ESL students in Victorian schools. Correlations (Spearman's rho) between all pairs of raters were

Melbourne Papers in Language Testing 1992 Volume 1.2 pp. 19-44. The Language Testing Research Centre, The University of Melbourne.

calculated, showing a range from -0.091 to 0.731. In general, the correlations were very modest, even for paired ESL teachers.

An earlier but more comprehensive study by Carlson et al. (1985), which led to the introduction of Test of Written English (TWE) by the Educational Testing Service (ETS), included a comparison of the holistic scores assigned to academic essays by 23 trained English teachers and 23 trained ESL teachers. The essays were written by both native and non-native speaker applicants for undergraduate and graduate higher education courses in the United States. Four writing samples were collected from 638 subjects. Each sample was read initially by two readers, one from each of the English and ESL teacher groups. Discrepancy marking was later employed in appropriate cases.

In the statistical analysis which followed, the data was organised in two different ways to examine the issue of inter-rater reliability. Firstly, for each of the four essay topics used in the study, the 638 original pairs of judgements (i.e. before eliminating any discrepant scores) were tabulated with score 1 as the first score and score 2 as the second score assigned. Score 1 could either be assigned by an English or ESL rater with score 2 then being from a rater in the other group. Inter-rater correlations (Pearson's r) were then obtained for each of the four topics, ranging from 0.66 to 0.74.

Secondly, the data was retabulated so that score 1 in each pair of judgements was the score assigned by the ESL rater and score 2 by the English rater. Carlson et al. (1985 : 61) argued that, if the ESL teachers assigned scores that were systematically higher or lower than the English teachers, then the recalculated inter-rater reliabilities may well have been higher than the original reliabilities. However, this was not the case. The mean scores assigned by the English and ESL raters were nearly identical and the interrater correlations (r = 0.67 - 0.72) were very similar to the first set of figures reported above. On the basis of these findings, Carlson et al. (1985 : 65) conclude that "the ratings of the English teachers and ESL teachers agreed very well".

In addition, the fact that the mean holistic scores for both groups were almost identical on each of the four essay topics used in the study suggests that English and ESL teachers do not significantly differ in the global ratings they assign to essays written by both native and non-native speaker students. Although appropriate tests of significance were not carried out on this data (t-tests, for example), it appears highly likely that the observed differences between the means in each of the four cases would be attributable to chance i.e. not significant.

This finding is confirmed in two more recent studies by Brown (1991) and Purpura (1992). In Brown's (1991) research project at the University of Hawaii (Manoa), designed primarily to investigate the relative writing abilities of native and non native speakers at the end of different first-year university composition courses (either English 100 or ESL 100), eight English and eight ESL faculty members rated 112 randomly assigned essays, 56 of which were written by English students and the other 56 by ESL students. Each essay was scored by two English raters and two ESL raters, with each rater marking 28 essays. A holistic six-point rating scale was used by all raters in marking the essays. The results indicated that there were no statistically significant mean differences between the ratings given by the English and ESL raters or between the native and non- native compositions.

Brown (1991) also examined the levels of agreement both between and within the two groups of raters. As in Morgan's (1990) study, low levels of correlation (although significant at $p < 0.05$ in all cases) were found <u>between</u> the two groups of raters (r = 0.36 - 0.58) and <u>within</u> both the English and ESL rater groups (r = 0.37 and 0.47 respectively) for all subjects. The combined reliability estimates (calculated by using an adaptation of the Kuder-Richardson formula 20) were more encouraging: 0.76 for all raters, 0.54 for the English raters and 0.64 for the ESL raters. Nevertheless, on the basis of these figures, it appears that the overall degree of inter-rater reliability in this study was not particularly strong.

In Purpura's (1992) study, 314 essays written by both English and ESL first-year undergraduate students were double marked by a group of 11 English and 6 ESL teachers at the University of California, Los Angeles (UCLA). Using the Rasch Item Response Theory program, FACETS (Linacre, 1988), again no significant differences between the global essay ratings of the two groups of teachers were found. This was true for both the whole group of subjects as well as English and ESL subjects considered separately.

In Brown's (1991) research project, markers were also asked to choose the best and worst features from among cohesion, content, mechanics, organisation, syntax and vocabulary of each composition in conjunction with assigning a global score. Analysis of the results showed that the two groups may have arrived at their global scores from somewhat different perspectives. In terms of positive features, overall English teachers were most influenced by syntax and cohesion in making their global assessments. Conversely, ESL teachers were most strongly governed by content followed by organisation. In terms of negative features, both groups of raters seemed to attend most to syntax with mechanics also being of importance to English teachers and content to ESL teachers (Brown 1991 : 601).

One significant limitation of this feature analysis, however, is that no results are given for English and ESL subjects taken separately. It is possible that the pattern of best and worst features for these two groups of subjects may differ considerably in either or both rater groups. For example, English teachers may attend to syntax as a positive feature more when rating the essays of ESL rather than English students. Conversely, ESL teachers may be more influenced by content, for instance, as a negative factor when marking ESL as opposed to English essays.

The results from Brown's (1991) feature analysis should be considered against the most common findings in the literature of direct writing assessment which indicate that raters are most strongly influenced by content and organisation in assigning holistic essay scores (Huot 1990a : 256). The only significantly different recent set of results emerges from a study carried out by Raforth and Rubin (1984). Using a design which involved the systematic manipulation of the quality of content and mechanics in a student essay as well as the type of intructions given to raters, they found that mechanics exerted a greater influence on markers' judgements than either content or rating instructions.

Although content and organisation generally appear to concern essay raters most in the literature on the factors influencing rater judgement about writing quality, it is important to note that most of this research has focused on native speaker essay samples only and generally does not appear to have included ESL specialists as raters.

One notable exception to this gap in the literature is a study by Mendelsohn and Cumming (1987) which examined the influence of language use (accuracy of syntax and morphology) and rhetorical organisation (clarity of overall structure) on the global judgements made by English, ESL and Engineering professors of essays written by first year university ESL students. The results indicated that were no differences in the relative importance attached by the three groups of raters to these two factors for the best or worst essays. However, for the middle range of ESL writing ability, it was found that ESL raters were more influenced by rhetorical organisation in making their judgements whereas the English raters did not seem biased in either direction. By contrast, the Engineering raters appeared to attribute more importance to language use in this instance.

It may well be that English and ESL teachers are influenced by different factors in assigning global scores to student essays. As suggested above, this may also depend on whether they are assessing the writing of native or non-native speakers.

3. Purpose

The central aim of this thesis was to investigate whether English and ESL teachers rate essays differently. The study involved, firstly, a comparison of the assessments made by four English and four ESL teachers of essays written by final year secondary English and ESL students. It also included an examination of the factors most strongly governing the global judgements of each of these two groups of raters.

The scoring procedure employed to explore these issues was both holistic (i.e. global) and analytical : essays were rated on both a global category and five analytical categories (see Section 5.1 for further detail). In calculating the total score for each essay the global category carried as much weight as the other five categories combined.

The study addressed the following research questions :

1. What level of agreement exists among all raters in relation to the total essay scores?

2. What level of agreement exists between the average total essay scores assigned by English and ESL teachers?

3. Is there a significant difference between the average total essay scores assigned by English and ESL teachers?

4. What level of agreement exists between the average essay scores assigned by English and ESL teachers on both the global and analytical scoring categories?

5. Is there a significant difference between the average global essay scores assigned by English and ESL teachers?

6. Which analytical categories (e.g. content, organisation or syntax) most strongly influence the global assessment of essays by English and ESL teachers taken separately?

4. The University of Melbourne Trial English Selection Test

This study is based on data gathered from the trialling in May 1992 of a written English test designed to assist with the process of selecting students for undergraduate study at the University of Melbourne. The test was developed in two versions for English and ESL students at the National Languages and Literacy Institute of Australia (NLLIA) Language Testing Centre, University of Melbourne.

In both versions of the test, candidates were given a choice of two reading passages accompanied by test tasks which included reading comprehension, evaluation of argument and an argumentative /persuasive essay. The two versions of the test were essentially the same except that the ESL version included a more extensive glossary of terms to assist reading comprehension, a limited amount of deletion in one of the reading passages and slightly reduced word length expectations for the essay.

The test was trialled on 484 final year secondary students in ten Victorian schools with a bias in favour of independent schools within metropolitan Melbourne. A fuller description and analysis of the trialling are given in the final report (O'Loughlin 1992).

A training program for markers drawn mainly from the participating schools but also from within the University was held shortly after the administration of the test. The secondary teachers were all experienced English and/or ESL specialists while the University assessors came from a variety of disciplines including Philosophy, Engineering and Fine Arts as well as English and ESL. Two parallel one-day training sessions were organised - one for raters of ESL papers and the other for raters of English papers. A common program was devised for both sessions focusing on the marking of sample scripts as a group and follow-up discussion of factors determining the assigning of scores to scripts. The same scoresheet was used to mark both English and ESL scripts in each of the sessions (see Appendix A: Scoresheet). The one salient difference between the two sessions was that English raters were trained using English sample scripts only while ESL raters were restricted to ESL sample scripts. The overall aim was to bring raters as close as possible to a consensus view of what they were assessing in each part of the test and of how to distinguish between levels of performance. All of the raters were then required to mark 35-40 scripts over a two week period.

In the marking process, in general, English scripts were double marked by assessors from the English rater group and ESL scripts by the ESL rater group. Final test results were calculated by averaging the total scores assigned by each pair of raters for each test candidate.

5. Method

5.1 Procedure

For the purpose of this study, 40 scripts (of which 20 were written by English students and the other 20 by ESL students) were set aside and independently assessed at the same time as the rest of the scripts were marked by a highly experienced group of four English and four ESL teachers drawn from the pool of trained raters. These eight raters were selected on the basis of their specialisation in either English or ESL (not both): all of them were native speakers of English, qualified teachers of English or ESL and had taught in secondary schools for more than five years.

Apart from the fact that, in all 40 cases, candidates had answered questions based on the same reading passage for the sake of comparability, they were randomly selected from the 10 participating schools. The 20 English and 20 ESL scripts were mixed together and then numbered from 1 to 40 before being photocopied into four complete sets of scripts. Each set was then divided so that 4 of the 8 raters read scripts 1-20 first and then exchanged them with another rater who had read scripts 21-40.

Prior to commencing the marking process, raters were informed that the researcher was interested in comparing how English and ESL students had performed in the test. However, the specific aims and details of the study were not revealed until all of the marking had been completed.

The final results for these scripts were later calculated by averaging the total test scores of the two most reliable English raters in the case of the 20 English scripts and the two most reliable ESL raters for the ESL scripts.

The focus of analysis in this study was the major component of the test - the argumentative essay. The reading passage which formed the basis of the essay presented an argument against extending the human lifespan through genetic engineering. The actual essay topic was "What do you think should be the average human lifespan?" As different word length expectations were given for the essay in the two versions of the test (400-700 for the English version and 300-500 for the ESL version), raters were instructed not to penalise any essays of more than 300 words in their scoring so as not to unfairly disadvantage ESL essays which totalled between 300-400 words.

The scoring method used for the essay, derived from McNamara (1990), was both holistic (i.e. global) and analytical. Raters were required to assign a score of between 1 and 6 (whole numbers only) on both a global category and five analytical categories : overall task fulfilment, arguments and evidence, organisation, appropriateness of language, control of linguistic features (grammar and cohesion) and control of presentation features (spelling and punctuation) (see Appendix A: Scoresheet). The global category (overall task fulfilment) carried as much weight as the other categories combined in calculating the total score for each essay. Raters were aware of this weighting when they were assessing the essays.

5.2 The intra-class correlation

The correlation statistic most commonly employed in this study was the intra-class correlation (Bartko, 1966). This statistic, unlike the more standard parametric Pearson correlation, provides a measure of actual agreement rather than simply linearity. Another difference is that it yields a single 'average' correlation co-efficient where more than two raters are used. It is computed by applying a one-way analysis of variance with each subject constituting a group. The intra-class correlation is then derived from the F-value using the following formula:

$$r_I = \frac{F - 1}{F + m - 1}$$

where m denotes the number of conditions for the independent variable i.e. the number of sets of ratings.

A P-value may be obtained for the correlation co-efficient using the appropriate F-value with (n-1, n(m-1)) degrees of freedom where n equals the number of subjects and m again represents the number of sets of ratings.

This type of correlation was employed following the example of a recent study by Elder (1992) comparing the ways subject specialists and ESL teachers construe the second language proficiency of non-English speaking background teacher trainees in secondary schools.

6. Results

The results for each of the six key research questions listed in Section 3 are reported below.

1. What level of agreement exists among all raters in relation to the total essay scores?

Table 1 below shows the intra-class correlations for the essay totals of all raters, English raters and ESL raters for all subjects as well as the English and ESL subjects considered separately. As explained in section 5.2.2, these figures represent average correlation co-efficients here as there were more than two raters in each case. In general, the amount of agreement between raters is quite low, even though the correlations were all significant at the 0.01 probability level.

	ALL RATERS (N=8)	ENG RATERS (N=4)	ESL RATERS (N=4)
ALL SUBJECTS (N=40)	0.56**	0.63**	0.53**
ENG SUBJECTS (N=20)	0.33**	0.45**	0.23**
ESL SUBJECTS (N=20)	0.58**	0.62**	0.56**

** p < 0.01 (two -tailed)

TABLE 1. INTER - RATER RELIABILITY
Intra-class correlations (r_I) for the essay totals of all raters.

Pearson correlations for all pairs of raters (all subjects only) were also calculated for the sake of comparison with the corresponding figures from the studies by Carlson et al. (1985) and Brown (1991) (see Tables 2, 3 and 4 below).

	R 1	R 2	R 3	R 4
R 1	1			
R 2	0.67**	1		
R 3	0.68**	0.70**	1	
R 4	0.53**	0.73**	0.62**	1

** $p < 0.01$ (two-tailed)

TABLE 2. INTRA - GROUP RELIABILITY
Pearson correlations (r) for the essay totals of English rater pairs (all subjects, N = 40).

	R 1	R 2	R 3	R 4
R 1	1			
R 2	0.55**	1		
R 3	0.60**	0.71**	1	
R 4	0.68**	0.61**	0.60**	1

** $p < 0.01$ (two-tailed)

TABLE 3. INTRA - GROUP RELIABILITY
Pearson correlations (r) for the essay totals of ESL rater pairs (all subjects, N = 40).

ENG	R1	R2	R3	R4
ESL				
R1	0.57**	0.55**	0.50**	0.70**
R2	0.72**	0.59**	0.69**	0.41**
R3	0.60**	0.79**	0.73**	0.64**
R4	0.68**	0.60**	0.56**	0.54**

** p < 0.01 (two-tailed)

TABLE 4. INTER - GROUP RELIABILITY
Pearson correlations (r) for the essay totals of English and ESL rater pairs (all subjects, N = 40).

While all of the correlations were significant at the 0.01 probability level, they are still rather modest overall. It is also worth noting that the intra-group figures (Tables 2 and 3) are very similar to the inter-group results (Table 4).

In relation to the earlier studies, these correlations are all consistently higher than the figures of Brown (1991 : 592) (r = 0.36 - 0.58 between the two groups of raters and 0.37 and 0.47 for within the English and ESL raters respectively, where N=112). On the other hand, the inter-group results are fairly much like those obtained by Carlson et al. (1985 : 62)
(r = 0.67 - 0.72 between English and ESL raters, where N = 638).

However, the comparison with the findings of Carlson (1985) and Brown (1991) should be regarded cautiously, firstly, because the differences in strength of correlation may be, in part, attributable to the variations in sample size and secondly, the results from those studies are based on holistic scores rather than total scores derived from adding together the holistic and analytical scores as is the case here. The more appropriate contrast is probably with the global category (overall task fulfilment) used in this study considered separately (see the discussion of the results for Question 4 below).

It should be stressed at this point that the rather modest intra-class correlations (as shown in Table 1) within the two groups of raters for their total essay scores does not necessarily invalidate comparisons between the rater groups in relation to either the total or categorical assessments. The reason for this is that the overall reliability estimates for each of the two groups of raters, computed by means of the Spearman-Brown Prophecy Formula (Henning 1987 : 83) using the Pearson correlations in Tables 2 and 3 above, are quite respectable - 0.88 for the English teachers and 0.87 for the ESL teachers. For the purpose of this study, this means that comparisons between the two groups, provided they rely on average scores, are quite legitimate. Any conclusions drawn from these comparisons, however, must be regarded with a degree of caution since individual differences between the four raters in each group are ironed out in the averaging process.

2. What levels of agreement exists between the average total essay scores assigned by English and ESL teachers?

Table 5 below shows the intra-class correlations (r_I) and Pearson correlations (r) for the average essay totals assigned by English and ESL teachers for all subjects and the two groups of subjects (English and ESL) taken separately. All of the figures were found to be significant at the 0.01 level. The Pearson correlations are included here for the sake of comparison with their intra-class equivalents.

SUBJECTS	N	r_I	r
ALL	40	0.73**	0.78**
ENGLISH	20	0.58**	0.65**
ESL	20	0.63**	0.72**

** p < 0.01 (two-tailed)

TABLE 5. INTER-GROUP RELIABILITY
Correlations (r_I and r) for the average essay totals assigned by English and ESL teachers.

Given, as previously noted, that the Pearson correlation is a measure of linearity only rather than actual agreement, it is not surprising that the relevant figures here are slightly higher than their intra-class equivalents. Although the Pearson statistic is not generally employed in the rest of this results section to calculate correlations, it is reasonable to assume that this pattern would be repeated in most other cases as well.

The Pearson correlation of 0.78 for all subjects suggests a reasonable amount of agreement overall between the the two rater groups. However, it should be noted that the corresponding intra-class correlation figure ($r_I = 0.73$), which is a truer measure of agreement, is the more accurate figure. The correlation results (both r and r_I) between the two groups of raters for English and ESL subjects when considered separately are more modest than for the whole group of subjects. This is at least partially due to the fact that the number of subjects was smaller in these two cases.

3. Is there a significant difference between the average total scores assigned by English and ESL teachers?

Two-tailed dependent t-tests on the differences between the average essay totals of English and ESL raters were carried out to examine this question. The results, as shown in Table 6 below, indicate that English teachers rated the whole group of English and ESL subjects significantly more harshly than ESL teachers. Furthermore, English teachers also rated ESL subjects

SUBJECTS	N	Mean Diff	t	P value
ALL	40	-0.76	-3.32	0.002**
ENGLISH	20	-0.60	-1.83	0.082 n.s.
ESL	20	-0.91	-2.838	0.011*

** $p < 0.01$ * $p < 0.05$ n.s. = not significant (two-tailed)

TABLE 6. t-tests on the differences between the average essay totals of English and ESL raters.

significantly more harshly than ESL teachers. This was almost true for English subjects as well.

It should be noted at this point that correlation statistics and t-tests examine different issues. This explains why the results in Table 5 (Question 2) and Table 6 here may appear, at first glance, to contradict each other. The t-test measures whether there is a significant difference between the means of two samples whereas correlation co-efficients provide a measure of the agreement (or, at least, linearity in the case of the Pearson correlation) between two (or more) data sets.

4. What level of agreement exists between the mean essay scores assigned by English and ESL teachers on both the global and analytical categories?

The intra-class correlation statistic was used to examine this issue (see Table 7 below). The correlation figures are derived from a comparison of the mean ratings allocated to candidates on each of the six criteria by the two groups of markers.

The intra-class correlations obtained for the global category (overall task fulfilment) provide an interesting comparison with their equivalents for the average total essay scores (see Table 5 above). While the figures for English subjects are almost the same, those for all subjects and, particularly, ESL subjects are higher. This finding will be further discussed in Section 7.

Overall, the level of agreement between the two groups of raters here is not particularly strong, although there is clearly greater agreement on the first three categories than the last three which, interestingly enough, are the criteria most directly concerned with actual language use. In addition, there is a higher level of agreement between the two groups of markers on four of the categories for ESL compared to English subjects, especially the global category.

	ALL Ss (N=40)	ENGSs (N=20)	ESL Ss (N=20)
Overall Task Fulfilment	0.77**	0.56**	0.80**
Arguments and Evidence	0.75**	0.60**	0.72**
Organisation	0.77**	0.70**	0.70**
Appropriateness of Language	0.51**	0.43*	0.27 n.s.
Grammar and Cohesion	0.62**	0.31 n.s.	0.51**
Spelling and Punctuation	0.56**	0.34 n.s.	0.50**

** $p < 0.01$ * $p < 0.05$ n.s. not significant (two-tailed)

TABLE 7. INTER-GROUP RELIABILITY
Intra-class correlations (r_I) between ratings assigned by English and ESL teachers for both global and analytical essay categories.

The Pearson correlation was also calculated for the global category (all subjects) to compare with the corresponding figure for the average essay scores as well as the findings based on holistic scores by Carlson et al. (1985) and Brown (1991). The result ($r = 0.78$ $p < 0.01$ two-tailed, $N = 40$) is identical to the one obtained for the average essay totals and higher than the relevant range of figures reported by both Carlson et al. (1985 : 62) ($r = 0.67 - 0.72$, $N = 638$) and Brown (1991 : 592) ($r = 0.36 - 0.58$, $N = 112$). However, these differences may be attributable, in part, to the differences in sample sizes as previously noted. In any case, this comparison should be treated cautiously since the correlation co-efficient here, unlike the other two studies, is based on averages rather than raw scores.

5. Is there a significant difference between the average global essay scores asssigned by English and ESL teachers?

Dependent t-tests were used to investigate this question (see Table 8 below). In each of the three cases (all, English and ESL subjects), the differences between the average global ratings were not significant. However, it is worth noting that the result for ESL subjects is almost significant at the 0.05 probability level.

SUBJECTS	N	Mean diff	t	P value
ALL	40	-0.17	-1.49	0.146 n.s.
ENGLISH	20	-0.09	-0.47	0.641 n.s.
ESL	20	-0.25	-1.86	0.079 n.s.

n.s. = not significant (two-tailed)

TABLE 8. t-tests on the differences between the average global essay scores of English and ESL raters.

These findings are at odds with those for the average essay totals (see Table 6, Question 3 above) but consistent with the results of Carlson et al. (1985), Brown (1990) and Purpura (1992) obtained using holistic scores. The comparison between these results for the global ratings is probably the more appropriate one to be made with the findings in these earlier studies. However, as in Question 4, it should be pointed out that the results here are based on averages rather than raw scores and therefore any conclusion based on this comparison with the other studies must be drawn tentatively. The discrepancy in this study between the results for the global ratings and those obtained for the average essay totals will be further discussed in Section 7.

6. Which analytical categories most strongly influence the global assessment of essays by English and ESL teachers taken separately?

A. ENGLISH TEACHERS

Table 9 below shows the intra-class correlations between the average ratings for the global category (overall task fulfilment) and the other five categories as assigned by English teachers.

	ALL Ss (N=40)	ENG Ss (N=20)	ESL Ss (N=20)
Arguments and Evidence	0.95**	0.92**	0.93**
Organisation	0.95**	0.93**	0.94**
Appropriateness of Language	0.89**	0.85**	0.96**
Grammar and Cohesion	0.85**	0.82**	0.75**
Spelling and Punctuation	0.87**	0.84**	0.82**

** $p < 0.01$ (two-tailed)

TABLE 9. ENGLISH RATERS : CORRELATION BETWEEN GLOBAL AND ANALYTICAL SCORES

Intra-class correlations (r_I) between the average scores on Overall task fulfilment and the other five categories for English raters.

Clearly there is very strong agreement between the global category and both content (arguments and evidence) and organisation. Content and organisation are very closely related themselves ($r_I = 0.94$ for all subjects $p < 0.01$) so it is difficult to determine which is contributing most to the global category.

Appropriateness is also strongly influencing the global category, especially for ESL subjects. In addition, organisation and appropriateness are very closely linked ($r_I = 0.92$ for all subjects $p < 0.01$) indicating that perceptions of appropriateness may influence judgements about essay structure and vice-versa.

In general, the correlations here are all high, suggesting that the judgements of English teachers on all of the analytical categories are contributing strongly towards their ratings on the global category. The correlations between grammar and cohesion and the global category are a little lower than the others, particularly for ESL students. It may be that English teachers feel less confident in making judgements about this category for second language learners and therefore it contributes a little less to their global judgements.

These findings can be compared with those of Brown (1991) who found that overall English teachers were most positively influenced by content and to a lesser degree by syntax and cohesion and most negatively by syntax and mechanics.

B. ESL TEACHERS

In the case of ESL teachers (see Table 10 below) there is also a very strong relationship between the global category and both content and organisation. Furthermore, again there is a high correlation between content and organisation ($r_I = 0.89$ for all subjects $p < 0.01$) making it difficult to determine which is contributing more to the global judgements of these raters.

While the other correlations between the analytical and global categories are weaker, it is worth noting that the criterion, grammar and cohesion is exerting a strong influence on the global category for ESL subjects but hardly any for English subjects. It may be that ESL teachers assume that most English students will be strong in this aspect of essay writing and therefore focus on it much less when assigning their global scores.

	ALL Ss (N=40)	ENGSs (N=20)	ESL Ss (N=20)
Arguments and Evidence	0.95**	0.90**	0.96**
Organisation	0.93**	0.80**	0.96**
Appropriateness of Language	0.54**	0.30 n.s.	0.47*
Grammar and Cohesion	0.73**	0.16 n.s.	0.84**
Spelling and Punctuation	0.62**	0.26 n.s.	0.65**

** $p < 0.01$ * $p < 0.05$ n.s. not significant (two-tailed)

TABLE 10. ESL RATERS : CORRELATION BETWEEN GLOBAL AND ANALYTICAL SCORES Intra-class correlations (r_I) between the average scores on Overall task fulfilment and the other five categories for ESL raters.

Grammar and cohesion is also fairly closely linked here with organisation for ESL subjects (r_I = 0.84) suggesting that perceptions of overall essay structure are influenced by judgements about grammar and cohesion and vice-versa. The ESL raters are also markedly more influenced by spelling and punctuation in assigning global scores to ESL rather than English essays. Generally, however, in making their overall judgements, ESL teachers appear to focus mainly on content and organisation for both groups of students and also on grammar and cohesion for ESL students only.

In Brown's (1991) study, on the other hand, ESL teachers were most positively influenced by content and, to a lesser extent, organisation. They were most negatively influenced by syntax and, to a lesser extent, content.

7. Discussion

7.1 Holistic versus analytical scoring

In general, the level of agreement between the English and ESL teachers (as measured by the intra-class correlation statistic) was higher for the global category than for the essay totals obtained by combining the global and analytical scores. Futhermore, while no significant difference was found between the global essay ratings of the two groups of teachers (as in the studies of Carlson et al. (1985), Brown (1991) and Purpura (1992)), this was not the case for the essay totals. The comparison of the average essay totals of the two rater groups indicated that overall, English teachers rated all of the essays significantly more harshly than ESL teachers. This was also true for ESL essays and almost for English essays taken separately.

Even though the global category carried as much weight as the five analytical categories in calculating the total scores, it is clear that these total scores, when averaged for each group of teachers, still reflected important differences in rater behaviour between the two groups. The analysis suggested, firstly, that the analytical categories were scored more harshly by English raters than ESL raters overall and, secondly, that the two groups of raters weighted these categories differently in arriving at their global judgements. While perceptions about content and organisation appeared to be the most dominant influences on both rater groups, English teachers were more strongly influenced by all of the other categories as well. The one other important category for ESL teachers was grammar and cohesion, but only for ESL essays.

These findings raise several issues in relation to scoring procedures. While the holistic method of scoring has consistently proved to be fairly reliable in the direct assessment of writing (Huot 1990b), it may well mask important differences between raters of different backgrounds and professional experience. The results in this study suggest that the analytical scoring method is more revealing about such dissimilarities between raters. The choice of scoring procedure, then, when different types of raters are used, is likely to determine whether or not these differences are highlighted and thus the overall level of inter-rater reliability.

The fact the holistic scoring method proved to be more reliable than the analytical scoring method in this study does not necessarily imply that it provides more valid assessments (i.e. truer, more accurate measures) of writing quality. Reliability is a necessary but not sufficient condition for test validity. On the basis of the findings in this research project, there can be no guarantee that the holistic scoring method used here actually allows for the real measurement of writing quality.

A great deal of research has been done on the reliability of the holistic scoring method, specifically inter-rater reliability. However, the focus on reliability appears to have been at the expense of an investigation of its validity. Huot (1990b : 204) underscores this point:

The most important side effect of the constant stress on reliability is that it has caused the profession to assume, confuse, and otherwise neglect the validity of holistic scoring procedures.

While this is certainly true, more research is needed on the validity of other scoring methods, including analytical scoring, as well as holistic scoring.

7.2 The assessment of ESL essays.

The comparison of the average essay totals indicated that English teachers marked ESL essays (taken separately) significantly more harshly than ESL teachers. This was almost the case for the global category as well. These findings have implications for school-based assessment of ESL written work carried out by English teachers in schools which are without an ESL teacher, particularly at the senior secondary level. In such instances, ESL students may well be disadvantaged compared to ESL students in other schools who are assessed by ESL teachers, particularly if the two groups of teachers do not participate jointly in group consensus marking or moderation on a regular basis (Morgan, 1990).

On the other hand, assessment of ESL essays by ESL teachers only is probably not desirable either. While it is obviously important that ESL teachers mark ESL writing, there is a risk that their scoring may not conform to the standards of the broader educational community. Hamp-Lyons (1991 : 326) shares this concern :

If only ESL teachers read ESL writing, the danger of losing sight of the expectations of the academy are very real.

This comment seems to be particularly pertinent to any future administration of the university selection test on which this study was based.

Where possible, therefore, ESL essays (and also English essays where comparability of scores between the two essay types is important) should be marked by an English specialist (or subject specialist if more appropriate) as well as an ESL specialist, especially if an analytical scoring method is used. Further research comparing the assessment of ESL essays by ESL teachers, English teachers and other subject-specialists using the analytical scoring method is needed.

8. Conclusion

While the results based on the global essay ratings of English and ESL teachers indicated that there was no significant difference between them, the findings based on firstly, the essay totals and, secondly, the category scores showed there were, in fact, important differences in their behaviour as raters both in relation to harshness of scoring and the factors influencing their global judgements. This suggested that the analytical scoring method may be more faithful to real dissimilarities which exist between raters of different backgrounds and professional experience than the holistic scoring method in the assessment of writing.

8.1 Further research

The suggestions for further research included in Section 7 are listed below:

1. How valid are the various scoring methods used in the direct assessment of writing?

2. How do the assessments of ESL essays by ESL, English and subject specialists compare with each other when an analytical scoring method is used?

3. Which has a stronger effect on reliability in the direct assessment of writing: the training process or the background and professional experience of raters?

References

Bartko, J.J. (1966) The intra-class correlation coefficient as a measure of reliability. *Psychol. Rep.*, 19 : 3-11.

Brown, J.D. (1991) Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, 25, 4 : 587-603.

Carlson, S., Bridgeman, B., Camp, R. and Waanders, J. (1985) *Relationship of admission test scores to writing performance of native and non native speakers of English.* (TOEFL Research Rep. 19) Princeton, N.J. : Educational Testing Service.

Elder, C. (1992) How do subject specialists construe language proficiency ? *Melbourne Papers in Language Testing*, 1,1 : 17-36.

Hamp-Lyons, L. (ed.) (1991) *Assessing second language writing in academic contexts.* Norwood, N.J. : Ablex.

Hatch, E. and Lazaraton, E. (1991) *The research manual : design and statistics for applied linguistics.* Rowley : M.A. : Newbury House.

Henning, G. (1987) *A guide to language testing : development, validation, research.* Cambridge. M.A. : Newbury House.

Huot, B. (1990a) The literature of direct writing assessment : major concerns and prevailing trends. *Review of Educational Research,* 60,2 : 237-263.

Huot, B. (1990b) Reliability, validity and holistic scoring : what we know and what we need to know. *College Composition and Communication,* 41, 3 : 201-213.

Linacre, J.M. (1988) *FACETS, a computer program for the analysis of multi-facted data,* Chicago : MESA Press.

McNamara, T.F. (1990) *Assessing the second language proficiency of health professionals.* PhD thesis, University of Melbourne.

Mendelsohn, D. and Cumming, A. (1987) Professors' ratings of language use and rhetorical organisation in ESL compositions. *TESL Canada Journal,* 5, 1 : 9-26.

Morgan, J. (1990) *ESL students and the new Victorian Certificate of Education common study.* M.A. thesis, University of Melbourne.

O'Loughlin, K.J. (1992) *Final report on the University of Melbourne Trial English Selection Test.* NLLIA Language Testing Centre, University of Melbourne.

Purpura, J.E. (1992) *Rater consistency between and among ESL teachers and writing program teachers.* Unpublished paper, Department of TESL/Applied Linguistics, University of Los Angeles, California.

Raforth, B.A. and Rubin, D.L. (1984) The impact of content and mechanics on judgements about writing quality. *Written Commmunication,* 1, 446-458.

APPENDIX A

UNIVERSITY OF MELBOURNE
TRIAL ENGLISH SELECTION TEST
SCORESHEET

Script Number ______________ Assessor ______________________________

English OR ESL (Please circle)

Option 1 OR Option 2 (Please circle)

Using the scales below, enter a number from 1 to 6 which best represents the candidate's performance for each category in the appropriate box on the right-hand side of the page. Use whole numbers only.

PART A TEXT COMPREHENSION

Complete 6 5 4 3 2 1 Incomplete ___

PART B EVALUATION OF ARGUMENT

Effective 6 5 4 3 2 1 Ineffective ___

PART C ARGUMENTATIVE / PERSUASIVE ESSAY

Note that the first category here ('Overall Task Fulfilment') carries as much weight as the other categories combined.

Overall Task Fulfilment
Excellent 6 5 4 3 2 1 Poor (i) ___

Arguments and Evidence
Well argued and supported 6 5 4 3 2 1 Poorly argued and supported (ii) ___

Organisation
Well organised 6 5 4 3 2 1 Poorly organised (iii) ___

Appropriateness of Language
Appropriate 6 5 4 3 2 1 Inappropriate (iv) ___

Control of Linguistic Features (Grammar and Cohesion)
Complete 6 5 4 3 2 1 Incomplete (v) ___

Control of Presentation Features (Spelling and Punctuation)
Complete 6 5 4 3 2 1 Incomplete (vi) ___

Signature ....

LANGUAGE TESTING CENTRE USE ONLY
Sub Total : Add (ii) - (vi) (vii) ___
Average (vii) / 5 (viii) ___
Essay Total : Add (i) and (viii) ___

TOTAL TEST SCORE
(Add Parts A,B and C) ___