Analytic rating scales: How diagnostic are they?

Kathryn Hill and Neomy Storch

1. Introduction

The advent of communicative approaches to language learning and teaching has in turn led to the use of more authentic assessment tasks. In the assessment of writing proficiency, there has been a move away from indirect tests, such as multiple-choice items, to direct tests, which measure writing proficiency through performance on a realistic task. Whereas indirect tests assess knowledge of correct forms, production of a sustained piece of writing enables candidates' ability to organise linguistic, semantic and schematic knowledge to be assessed (Mullen, 1980: 161).

However, an issue which arises is how performance on direct writing tests can be measured most effectively. The main approaches to the performance assessment of writing may be characterised as either 'objective' or 'direct'.

Objective assessment procedures usually involve frequency counts of particular linguistic features of the text, such as errors, length and complexity of t-units, cohesive devices etc. For example, marks may be awarded for the presence of a cohesive device and deducted for syntactical errors. Whilst this type of scoring instrument minimises the amount of subjective judgement involved in rating, it may have questionable validity as a method of assessing writing proficiency (Perkins, 1983) as it does not consider the effectiveness of the communication as a whole. So, rather than concentrating on isolated aspects of production, e.g. grammatical accuracy, there is a move to have raters consider other qualities of written production (Shohamy, 1985).

There are three types of 'direct' scoring methods: holistic, analytic and primary trait. With holistic or global scoring methods, one or more raters award an overall proficiency score based on their general impression of the performance (Perkins, 1983). The primary trait method focusses on a particular attribute of performance considered most salient to the nature of the rhetorical task. In an argumentative essay, for example, the trait might be quality and

Melbourne Papers in Language Testing 1994 Volume 3.1 pp. 50-65. The Language Testing Research Centre, The University of Melbourne.

quantity of ideas and evidence (Perkins, 1983). The third direct method, analytic (or multi-trait) scoring, breaks writing down into its various component parts (for example organisation, ideas, control over syntactical features and vocabulary) for the purposes of scoring. The total score for the piece of writing may be either an aggregate or an average of the component scores. In some cases, the final score is a combination of both analytically and holistically derived scores.

One of the advantages claimed for analytic scoring instruments is that, because they permit performance on different facets of writing to be assessed and reported, they can be used to provide valuable diagnostic information (Hamp-Lyons, 1991). Hamp-Lyons argues that most candidates do not have a "flat" profile, that is, they do not score equally well on all criteria, and that an aggregate or averaged score may disguise a marked weakness on a particular criterion. She suggests that reporting scores on each of the analytic criterion:

"...offers the potential for providing information that can be used in language instruction programs for making fine-grained initial placements or needs diagnosis. A writer whose score information suggests she is weak on syntactic structures but strong in vocabulary might be placed in a grammar class as well as at the appropriate level of the writing course sequence; another writer, whose score information suggests he has strong grammatical skills but has little of substance to say may be placed in a reading course..." (Hamp-Lyons, 1991: 242).

Following this suggestion, this study sets out to examine the effectiveness of using an analytic scoring procedure to provide diagnostic information based on performance on a direct writing test.

Measurement of performance on any test needs to be both valid and reliable. Validity "... refers to the degree to which the evidence supports the inferences that are made from [test] scores. " (American Psychological Association, 1985: 9). By 'reliability' we mean the extent to which an instrument produces consistent and accurate results each time the test is administered (Hatch & Lazaraton, 1991).

According to Mullen (1980) there are two sources of unreliability in performance assessment of writing: the topic and the judges. In this study we are interested in the second source, that is, the raters. According to McNamara & Adams, one of the greatest threats to the reliability of performance tests is the increased subjectivity of assessment. This is because "raters contribute an additional source of variation to the measurement (additional, that is, to the variation associated with test items)" (1991: 1). For example, raters may tend to be harsher or more lenient relative to each other, or totally inconsistent in their scoring.

To improve reliability of subjectively scored tests it is recommended, inter alia, that each performance be assessed by more than one rater and that raters be highly trained (Perkins, 1983). However, rater training alone cannot guarantee reliability and inter-rater reliability needs to be investigated routinely. Furthermore, whilst investigations of inter-rater reliability usually focus exclusively on the final, aggregate scores for diagnostic purposes, the reliability of scoring on individual analytic criteria is also important. Hamp-Lyons and Henning(1991: 362), in a study investigating the use of an analytic scoring procedure to assess the writing performance of adult non-native English speakers, found interrater reliabilities (for three raters) were not high for individual criteria, with reliabilities ranging from 0.608 to 0.905. If raters are not rating reliably on each of the criteria, we can have little confidence in the diagnostic information these criteria are intended to convey.

Furthermore, although raters may be rating very consistently to the extent that a score from one rater predicts reliably the score given to the same subjects by another rater, they may not necessarily be awarding the same scores. That is whilst, the two raters may agree in their overall ranking of candidates, one rater may be consistently harsher than the other. Therefore, it needs to be established whether any differences in scores between raters are significant.

One claim made for analytic scoring is that it makes assessment more objective and hence more reliable because, in theory, it forces examiners to be more explicit about the assumptions underlying their assessment (Weir, 1990: 63; Perkins, 1983: 655). What is of interest for the purposes of this study is whether raters are actually using the analytic criteria in the way that was intended or whether they are applying some other measure of performance. McNamara (1990) posits the existence of a "deep seated" rater orientation to certain features of writing which may not be explicit in the scoring criteria. Moreover, this orientation seems to withstand the effect of rater training.

A study comparing marking of academic essays by ESL teachers, English teachers and university academic staff found that raters seem to share a general, often unspoken, agreement, concerning the expected standards of academic writing competence (Carlson & Camp, 1985 cited in McNamara, 1990). Researchers have found that these 'standards' often relate to levels of grammatical accuracy (McNamara, 1990), though, two studies of the assessment of University ESL students' essays using an analytic method, found that ratings on vocabulary usage accounted for most of the variance in the overall score (Mullen,1980; Astika, 1993).

The underlying assumption of an analytic scoring instrument is that each of the analytic criteria are assessed independently and that each contributes equally to the total score (though, in some cases, the test designers may decide to weight certain criteria differently). It is this assumption that allows for the adding and averaging of the scores to yield the final, overall score. Hence, a score given on one criterion, such as grammar or vocabulary, should not influence scores given on other criteria, nor contribute unequally to the final, overall score . If a single criterion appears to be strongly related to scores on all other criteria and to the overall score, this may indicate that the instrument is not being used as intended.

It is important to remember that the analytic criteria selected for a test reflect what are considered by the test designers to be important and measurable aspects of writing performance. If raters are not actually using the analytic criteria in the way that was intended but instead are applying some other measure of performance, i.e. their own internalised constructs, this poses a threat to the construct validity of the test (McNamara, in prep.).

In summary, a test score can be seen as the result of an interaction between candidate ability, the writing stimulus, the rater and the scoring method. In this study we are interested in the last two variables, specifically the reliability of ratings and the validity and reliability of using an analytic scoring instrument for diagnostic purposes. The hypotheses we wish to investigate in this study are as follows:

1. H_0 — there is no significant relationship between the scores of pairs of raters on each of the four analytic criteria and the overall score.

2. H_0 — there is no significant difference between the scores of pairs of raters on each of the four analytic criteria or on the overall score.

3. H_0 — there is no significant relationship between the four analytic criteria with each other or with the overall score.

2. Method

2.1 Subjects

The data for this study comes from the 130 candidates who sat the test in January and February, i.e. before the commencement of the 1993 academic year. Candidates, who self-selected for the test, came from a wide range of language backgrounds, the majority being Asian (L1 backgrounds included Indonesian, Malay, Vietnamese, Japanese and dialects of Chinese) and then European. There were roughly equal numbers of overseas and resident students. Approximately three quarters of the cohort were enrolled in undergraduate courses in a range of faculties, including Engineering, Commerce and Arts with the balance enrolled in post-graduate courses.

2.2 Procedure

The ESL test was developed by the ESL Program staff in November, 1992. The main purpose of the test is diagnostic. It is used to identify those students already enrolled at the university, either as undergraduate, postgraduate (including visiting scholars) and continuing education students who may require concurrent English language support in the form of lunch-time skills classes or individual tutorials. Due to limited places in these classes, students are only allowed to take two per semester and hence need guidance as to which classes they should take. This advice is based on test scores (including diagnostic information provided by individual analytic scores on the writing sub-test) and the particular language demands of the student's intended course of studies.

The test consists of three subtests: Listening and Note-taking; Reading Comprehension; and Writing. The three subtests are integrated in terms of the materials used and skills tested. All are based on the same general topic and students' listening notes and reading text are retained to be used as the context from which to draw ideas for the writing task. The writing task consists of a 200 – 300 word argumentative-style essay on a choice of two topics.

The writing sub-test uses an analytic scoring instrument. In this system, performance is assessed on each of four analytic criteria (Communicative Quality, Arguments, Grammar and Vocabulary) on a scale of 0 to 9. A final, 'Overall' score is derived by averaging the four analytic scores. This instrument differs from other analytic instruments in that performance at each point on the scale is described. (As these descriptors are confidential, no copy has been attached).

The raters were all experienced ESL teachers and were experienced in using analytic scoring instruments. Before rating commenced, raters underwent further training on trial papers, to ensure a common interpretation of the scoring instrument. Papers were randomly assigned to randomly paired raters. Seven raters formed five pairs, both members of each pair independently rating roughly 30 essays (see Table 1). Each essay was double marked. Data from one pair of raters was incomplete and was therefore not included in this study. There was no overlap of papers between different pairs of raters. Individual pairs conferred after scoring was completed and, if their initial ratings disagreed, arrived at a single, 'agreed' overall score for each essay. This conferring was only done for the final, overall score.

	Raters
Pair 1	Rater 'H' & Rater 'J'
Pair 2	Rater 'R' & Rater 'J'
Pair 3	Rater 'H' & Rater 'R'
Pair 4	Rater 'N' & Rater 'C'

Table 1:

For reasons of expediency (and because it was assumed, prior to this study, that the ratings were reliable) marks were reported showing the analytic scores given by a single (i.e the first) rater and the overall 'agreed' score (i.e. from both raters).

2.3 Analysis

1. Inter-rater reliability

H_0 — there is no significant relationship between the scores of pairs of raters on each of the four analytic criteria and the overall score.

Interrater reliability was estimated using Pearson's Product-moment correlation. To establish that the assumptions of this procedure had been met we first investigated whether the data was continuous. There are several precedents for treating what is ostensibly ordinal data as interval data (McNamara, 1990; Mullen, 1980). According to Hatch & Lazaraton (1991: 179) interval scores can be summed to give a single numerical value that reflects each individual's proficiency, which is exactly how scoring procedures such as the IELTS1 bandscales are used. In this case scores between 0 and 9 were awarded for each of the four analytic criteria to give a maximum possible aggregate of 36. The final (or Overall) score is an average of the four scores. This procedure assumes that bandscales are equi-distant although, as Pollitt & Hutchinson point out, this assumption is generally not tested (1987: 74). In other words, it is possible that the distance between a score of 3 and a score of 4, for

example, may not be the same as that between 4 and 5. Hence, to be confident that our correlations were not distorted by the use of data that may not be truly interval we also calculated a Fisher-Z transformation for ordinal data (Hatch & Lazaraton, 1991: 533) (Table 2).

The 0.05 significance level was set for this test.

	Pearson's Correlations			Fisher Z Transformation			T-Tests				
	r	r^2	p	z	r	r^2	p	paired t	p	Wilcoxon's	p
Pair 1											
Comm.	0.562	0.32	<0.01	0.612	0.54	0.29	<0.01	-2.04	0.05	18	0.039
Arg	0.566	0.32	<0.01	0.786	0.66	0.44	<0.001	-0.42	0.68	39.5	0.701
Vocab	0.627	0.39	<0.01	0.851	0.69	0.48	<0.001	2.2	0.38	93	0.065
Grammar	0.636	0.40	<0.01	0.862	0.7	0.49	<0.001	0.42	0.68	88.5	0.586
Overall	0.711	0.51	<0.001	0.94	0.74	0.55	<0.01	0.94	0.36	50.5	0.0388
n=24, df=22								n=24, df=23			
Pair 2											
Comm.	0.801	0.64	<0.001	1.047	0.78	0.61	<0.001	1.16	0.26	N/A	
Arg	0.832	0.69	<0.001	1.085	0.8	0.64	<0.001	-1.16	0.26		
Vocab	0.883	0.78	<0.001	1.158	0.82	0.67	<0.001	1.28	0.21		
Grammar	0.908	0.82	<0.001	1.208	0.83	0.69	<0.001	1.67	0.11		
Overall	0.952	0.91	<0.001	1.294	0.86	0.74	<0.001	-0.7	0.49		
n=22, df=20								n=22, df=21			
Pair 3											
Comm.	0.667	0.44	<0.001	0.895	0.72	0.52	<0.001	-2.21	0.038	26	0.057
Arg	0.632	0.40	<0.001	0.85	0.69	0.48	<0.001	-0.94	0.36	45	0.41
Vocab	0.707	0.50	<0.001	0.94	0.74	0.55	<0.001	-2.11	0.047	19.50	0.075
Grammar	0.756	0.57	<0.001	0.998	0.76	0.58	<0.001	-1.02	0.32	49.5	0.352
Overall	0.792	0.63	<0.001	1.034	0.78	0.61	<0.001	-0.49	0.63	23	0.683
n=23, df=21								n=23, df=22			
Pair 4											
Comm.	0.619	0.38	<0.001	0.84	0.69	0.48	<0.001	1.37	0.18	N/A	
Arg	0.726	0.53	<0.001	0.96	0.75	0.56	<0.001	1.14	0.26		
Vocab	0.774	0.60	<0.001	1.001	0.76	0.58	<0.001	2.24	0.026		
Grammar	0.828	0.69	<0.001	1.086	0.79	0.62	<0.001	2	0.05		
Overall	0.84	0.71	<0.001	1.099	0.8	0.64	<0.001	1.65	0.11		
n=30, df=28								n=30, df=29			

Table 2:

2. Significance of difference between scores

H_0 — there is no significant difference between the scores of pairs of raters on each of the four analytic criteria or on the overall score.

A dependent t-test was used to test this hypothesis. However, a problem emerged with the score distribution for one rater, Rater 'H', who tended to use only the higher end of the rating scale. To get around this problem it was decided to use the Wilcoxon Matched Pairs Signed-Ranks test, as well as the dependent t-test, for this part of the data to see if results remained significant. Significance was set at 0.05 for a two-tailed test.

3. The scoring instrument

H_0 — there is no significant relationship between the four analytic criteria with each other or with the overall score.

As with inter-rater reliability, a Pearson's correlation was performed for this test. Fifty sets of scores were randomly selected. Although individual, rather than 'agreed' scores were used for these correlations, no correction was made for rater reliability as this would involve a separate adjustment for each set of scores (reliability can be only calculated in a single measure if all raters rate all subjects). As the overall score contains the values of each of the subscores, corrections were performed for part-to-whole correlations to ensure independence of data (Hatch & Lazaraton, 1991: 437). Significance was set at 0.05 for a two-tailed test.

3. Results & Discussion

All the correlations for inter-rater reliability for each of the analytic criteria and the overall score were significant at the 0.05 level. As there was little difference between the Pearson's r and the value for r after the Fisher Z Transformation (Table 2) only Pearson's r will be reported. All four pairs of raters achieved the highest correlation on overall scores, with reliabilities ranging from r=0.71 to 0.95. However, the correlation for Pair 1 (r=0.71) is considered unacceptably low. r 2 estimates how much of the variance in one measure can be accounted for by the other (Hatch & Lazaraton, 1991: 441). In other words, it indicates the accuracy of the prediction when the score assigned by one judge in a pair is used to predict the score given by the other (Mullen, 1980: 165). Hatch & Lazaraton recommend an r value of between 0.8 and 1 to demonstrate that the two raters are measuring the same thing (1991: 441). By these standards none of the correlations for Pair 1 were acceptable.

As explained earlier, whilst most studies are only interested in correlations on the final score, this study is also concerned with reliability on each of the analytic criteria as well. As can be seen in Table 2, the range of reliabilities, from r=0.562 to 0.908, were even less satisfactory than in the Hamp-Lyons and Henning study, where reliabilities ranged from 0.608 to 0.905 (1991: 362). For all pairs the correlations followed a fixed hierarchy of strengths of relationship. The highest correlations after Overall were on Grammar followed by Vocabulary, Argument (except pair 3) and lastly, Communicative Quality (except pair 3). However, only Pair 2 demonstrated an acceptable level of inter-rater reliability on all four analytic criteria (Grammar r = 0.908; Vocabulary r = 0.883; Argument r = 0.832 and Communicative Quality r = 0.801).

That raters demonstrated the greatest agreement on Grammar and Vocabulary is not surprising when considering that, of the four analytic criteria, they are the most 'concrete' and therefore perhaps the most easily applied. Communicative Quality, which appears to involve the most 'subjective' judgement, is also the criterion with the lowest correlations. A comparison of the descriptor for Grammar "...intrusive subject/verb and tense agreement errors occur..." with Communicative Quality, "This is a satisfactory essay..." clearly demonstrates this objective/subjective dichotomy. In other words, the results suggest that the more 'subjective' the criterion, the less reliably it is applied.

Dependent t-tests were used in conjunction with the correlations to determine if there were significant differences in the range of scores given by individuals within each pair (Table 2). Significant differences were found in the ratings from Pair 4 on Grammar (paired t=2.24, df=29, p=0.026) and Overall (paired t=2, df=29, p=0.05). This was unexpected because this is where they achieved their highest correlations (Grammar: r=0.83, df=28, p<0.001; Overall: r=0.84, df=28, p<0.001). What this means is that, although their scores are closely parallel, they are not equivalent in terms of harshness. In other words, one of the raters is a more lenient judge than the other. As there were no significant differences for Pair 2, who demonstrated high reliability on all measures, it can be concluded that raters within this pair were both interpreting the four criteria and awarding individual scores in the same way. Differences in scores for Argument and Grammar for both Pairs 1 and 3 were significant for the dependent t-test and were close to significant for the Wilcoxon Matched-Pairs Signed-Ranks test. As both pairs demonstrated poor reliability overall it appears that they are inconsistent with each other in interpreting the criteria, Argument and Grammar, across all subjects.

There is reason to believe that Rater 'H', who appears in both Pair 1 and Pair 3, is the greatest single source of unreliability. As discussed earlier, her data is skewed by a tendency to use only the higher end of the rating scale. Furthermore, whilst Rater 'J' rates highly reliably in Pair 2, together with Rater 'H' in Pair 1 she achieves the lowest reliabilities of the group. If Rater 'H' could be singled out for further training or removed from rating altogether, reliabilities may be expected to improve significantly.

Looking at the scoring instrument itself, correlations were performed for each of the four analytic criteria with each other and with the overall score (corrected for part-to-whole correlations) (Table 3). Once again we found that the highest correlations were obtained for Grammar followed by Communicative Quality, Vocabulary and Arguments. According to our results a score on Grammar is the best predictor of the Overall score, accounting for 0.84 of the variance. The Grammar score is also the best predictor of scores on Communicative Quality and Vocabulary. Looking first of all at the relationship between Grammar and Communicative Quality, our findings support what Politzer & McGroarty (1983) refer to as a "...general tendency for high linguistic competence to correlate with high communicative competence..."(626). As far as Grammar and Vocabulary are concerned, a high degree of overlap is probably inevitable given that problems with vocabulary (e.g. word formation) are often also grammar problems.

r

	Comm	Arg	Vocab	Grammar
Arg	0.789			
Vocab	0.814	0.675		
Grammar	0.878	0.77	0.869	
Overall*	0.899	0.786	0.844	0.915

*Adusted for part-to-whole correlations

r^2

	Comm	Arg	Vocab	Grammar
Arg	0.62			
Vocab	0.66	0.45		
Grammar	0.77	0.59	0.76	
Overall*	0.8	0.62	0.71	0.84

*Adusted for part-to-whole correlations

n=99 df=97, $p<0.001$

Table 3:

The results indicate that Communicative Quality is a slightly better predictor of a score on Argument than Grammar. Again, this is not surprising when considering that organisation of ideas and use of evidence involve non-linguistic skills and are more related to communication than to grammatical accuracy, i.e. Grammar and Argument are measuring something qualitatively different.

The next step would be to undertake a stepwise multiple regression analysis with the overall score as the dependent variable to determine the relative contribution of the four analytic criteria to the variance in the overall score. For the time being though, the pattern which emerges from these results is that not only is the highest inter-rater reliability achieved for Grammar, but that Grammar correlates most strongly with the three other criteria and the overall score.

If further analysis confirms that scores are being most strongly influenced by Grammar, the next question is which particular aspects of grammar are involved? Studies attempting to determine which feature, or combination of objectively measured features, might discriminate most highly among holistic evaluations have produced conflicting results (Huot, 1990). However, a number of researchers have found that objective measures, such as T-unit length, discriminated more reliably in assessment of more advanced candidates than with weaker candidates (Homburg,1984; Flahive and Snow, 1980). Therefore, it would be useful to compare the influence of the Grammar subscore at high and low levels of writing proficiency before attempting to determine precisely which aspects of grammar raters were attending to.

It was suggested earlier that the highest inter-rater reliabilities were achieved for Grammar and Vocabulary because they are 'concrete' (compared with Communicative Quality and Argument) and therefore easier to apply. To address this problem, some work perhaps needs to be done on rewriting the descriptors to improve rater agreement on what the more subjective analytic criteria mean in relation to writing performances. However, disagreement on how to apply the criteria may be symptomatic of a lack of real agreement about the qualities of good writing (apart from grammar) in the first place (Perkins, 1983: 654). This brings us back to the question of the construct validity of the scoring instrument. Essay writing is an integrated skill and the division of the performance into four discrete analytic criteria in a way contradicts this belief. Furthermore, it may not be simply that grammar is more 'concrete' than the other criteria, but that grammar is at the "core of language learning and more powerful in terms of generalizability than any other language features" (Davies, 1978, in Weir, 1990: 18).

To conclude, the analytic rating scales do not appear to be reliable as a diagnostic tool as they are presently used. Despite the use of expert trained raters and double marking of all papers, three of the four analytic criteria were not being interpreted with acceptable levels of reliability. If diagnostic information is to be meaningful, the results of this study suggest that 'agreed' scores need to be derived for each of the analytic criteria as well as for the overall score. In addition, rater reliability within, and between pairs needs to be assessed routinely and scores adjusted accordingly. However, the use of multiple raters assumes that individual raters are equally consistent (Huot, 1990) and, as this study demonstrated, this is not always the case. Ongoing monitoring and rater training is therefore essential. Finally, it is unclear how the unintended influence of Grammar on the other scores can be redressed. Whilst it is obviously a threat to construct validity, this phenomenon is by no means unique to this study.

References

Astika Bush Gede (1993) Analytical Assessments of foreign students' writing. RELC Journal, Vol. 24, No. 1, pp 32 – 60.

Brown, J. D. and K. M. Bailey (1984) A categorical instrument for scoring second language writing skills. Language Learning, vol 34, 4, pp 21 – 42.

Celce-Murcia, M. (1990) Discourse analysis and grammar instruction. Annual Review of Applied Linguistics, 11 pp 135 – 151.

Dickins, Pauline M. Rea and E. G. Woods (1988) Some criteria for the development of communicative grammar tasks. TESOL Quarterly, Vol. 22, No. 4, pp 623 – 651.

Evolva, J., E. Mamer and B. Lentz (1980) Discrete point versus global scoring for cohesive devices. In J. W. Oller Jr and K. Perkins (ed) Research in language testing. Rowley Massachusetts: Newbury House, Chap. 16.

Flahive, D. E. and B. G. Snow (1980) Measures of syntactic complexity in evaluating ESL compositions, In J. W. Oller Jr and K. Perkins (ed) Research in language testing. Rowley Massachusetts: Newbury House.

Hamp-Lyons, L. (1991) Scoring procedures for ESL contexts. In L. Hamp-Lyons (ed) Assessing second language writing in academic contexts. New Jersey: Ablex. pp 241 – 276.

Hamp-Lyons, L. and G. Henning (1991) Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. Language Learning, Vol. 43, No. 3, pp 337 – 373.

Hatch, E. and A. Lazaraton (1991) The research manual. Research and statistics for Applied Linguistics. Rowley, Mass.: Newbury House.

Henning, G. (1987) A guide to language testing. Cambridge, Mass.: Newbury House.

Homburg, T. J. (1984) Holistic evaluation of ESL compositions: can it be validated objectively? TESOL Quarterly 18, 1: 87 – 107.

Hughes, A. (1989) Testing for language teachers. Cambridge: Cambridge University Press.

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

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

McNamara, T. F. (in preparation) Second language performance testing theory and research. New York: Longman.

McNamara, T. F. and R. J. Adams (1991) Exploring rater behaviour with Rasch techniques. ERIC Document Reproduction Service, No. ED 345498.

Mullen, K. A. (1980) Evaluating writing proficiency in ESL. In J. W. Oller and K. Perkins (ed) Research in language testing. Rowley, Mass.: Newbury House, pp 160 – 170.

Mullineaux, A. (1981) Handbook of experimental design and statistics. Department of Applied Linguistics: Birkbeck College, University of London.

Pollitt, A. and C. Hutchinson (1987) Calibrated graded assessments: Rasch partial credit analysis of performance in writing. Language Testing 4, 1: 72 – 92.

Shohamy, E. (1985) Experimental Edition. A Practical Handbook in Language Testing for the Second Language Teacher, Shakad, Ramat Aviv, Israel.

Weir, C. J. (1990) Communicative language testing. New York: Prentice Hall.

1 International English Language Testing System — a 4 skill test of academic English

Melbourne Papers in Language Testing

I wish to subscribe to the Melbourne Papers in Language Testing. I enclose a cheque for the following:

Subscription (Vol 3,2 and Vol 4,1)	(Aus. $16.00)	□	*If only 1 issue required please circle which one and halve the costs of subscription and postage*
Air mail postage (overseas subscribers)	(Aus. $9.00)	□	

Name (Please print) ....

Address ....

....

....

Please make cheque payable to: NLLIA Language Testing Research Centre and send it with completed form to:

NLLIA Language Testing Research Centre, Department of Applied Linguistics and Language Studies, The
University of Melbourne, Parkville, Victoria 3052, Australia