The development and validation of a writing test and an analytic scoring scheme used in the ESL program, the University of Melbourne

Neomy Storch

1. INTRODUCTION

The advent, in the past 20 years, of the communicative paradigm in second language instruction has had a major impact on language testing. There has been a shift to direct, performance-based tests which assess the learners' ability to use the second language effectively in situations and on tasks relevant to the learner (Brindley, 1986:1). Direct assessment has become the most common means of evaluating writing yet research in the area has shown that great care needs to be taken in the design of direct tests and of the procedures used to assess learners' performance on such tests (Huot, 1990).

Recently, two staff members from the ESL Program at Melbourne University undertook to redesign the format of the writing test used in the assessment of students in the Advanced English as a Second Language (AESL) course and at the same time devise a suitable scoring instrument which would provide reliable and relevant information about students' achievement in the course.

This paper documents the development of the test and of the scoring scheme. It begins by giving brief background information about the AESL course and its assessment procedures to date and proceeds to explain the reasons for the changes. It then evaluates and validates the test and the scoring scheme using traditional and IRT data analysis as well as feedback from students and fellow staff in the ESL Program.

Melbourne Papers in Language Testing 1993 Volume 2.2 pp. 1-34. The Language Testing Research Centre, The University of Melbourne.

2. BACKGROUND INFORMATION

2.1. The AESL Course and its assessment procedures to date

The Advanced English as a Second Language (AESL) is a credit bearing EAP (English for Academic Purposes) course based on a content-skills curriculum. The content syllabus focuses on themes in contemporary Australian society and its origins; the skills syllabus is based on the development and consolidation of academic language skills; that is, skills required for understanding lectures, reading academic texts effectively and writing academic assignments.

Students come from a number of faculties (Arts, Law, Education, Engineering, Agriculture, Science and Medicine) and from a range of language backgrounds, although the predominant language groups are Chinese and Japanese. Both graduates and undergraduates are accepted. Students' language proficiency is assessed prior to entry to the course. The entrance exam consists of reading, listening and writing sub-tests similar in format to the exam used at the end of each semester. Students' writing proficiency, an important consideration in admission to the course, needs to be in the range of 5 - 7 on the IELTS band scales upon entry to the course.

Students' final grade for the course is a composite mark derived from marks on continuous assessment requirements, two major assignments and two end-of-semester exams. These are multi-skill, integrated exams composed of reading, listening and writing sub-tests all dealing with the same or similar issues. To date students were required to use their notes from the listening sub-test and the text from the reading sub-test as input for their final test: the writing sub-test. In this writing test, students were given a choice of 2 topics and were required to write within an hour one essay of approximately 500-750 words (about 1.5 - 2 pages).

The writing test is a direct test; testing students' performance on a 'realistic', representative task. That is, students are required to demonstrate their academic writing proficiency: their ability to marshal their linguistic, semantic and schematic knowledge in

extended, organised prose on a task which simulates the type of tasks students will encounter in the university setting (Shih, 1986; Greenberg, 1986). It serves largely as an achievement test but in view of the nature of the course it is also an academic proficiency test; that is, the test aims to assess the extent to which students have mastered the skills dealt with and practised in the AESL writing classes and which have been identified as important across many academic disciplines and courses (Johns, 1986). The topic and discourse mode chosen for each writing test is based on work done in the preceding semester's reading and writing classes.

The writing test was assessed using a slightly modified version of the IELTS writing band scales, these being internationally accepted and recognised measures of academic language proficiency. Each test was double marked, the overall grade being a 'negotiated' or averaged score.

2.2. Concerns about the AESL writing test and its assessment

Although staff were generally pleased with the integrative nature of the exams and their performance orientation, there has been mounting dissatisfaction with both the writing test format and the marking scheme.

The main concern with the test itself was whether it was able to test students' ability to analyse and synthesise a range of sources into an appropriate rhetorical form. This is a skill on which a large portion of class time was spent commensurate with its importance in a university context (Johns, 1986). It had been observed that students tended to over-rely on the text given in the reading sub-test (usually only one text was given) and made very little attempt to utilise their notes from the listening sub-test. Furthermore, in the past students complained that they ran out of time and hence could not complete the essay.

In terms of the assessment procedures, as Hamp-Lyons (1991) points out, a rating scheme needs to have input from the Program in which it is used. It needs to reflect the needs which the teachers in the program have identified as important in their own context and which suits their specific group of learners and

purposes. It was generally felt by staff that the currently used scheme did not provide for the assessment of students' synthesising skills nor for their ability to acknowledge sources used appropriately. These skills, as mentioned above, were considered to be very important in this context and hence needed to be assessed. Thus it was decided to change the format of the writing test and to design a new assessment scheme.

3. THE NEW AESL WRITING TEST

Literature in language testing has shown that a number of factors need to be taken into consideration in the design of writing tests. Factors such as the number of tasks, the topics, the discourse mode called for, the wording and structure of the rubric and the time given to complete the test may all influence the nature of the writing and in turn have consequences for test results.

It was decided that the new test should be longer (90 minutes) and provide students with stimulus materials in the form of short texts dealing with a range of views related to the essay topic. The reading and listening sub-tests were to be maintained in their original format thus providing students with additional input.

Despite the recommendations made by a number of writers in language testing (Carroll, 1980; Huot, 1990) that a number of tasks be set and the longer duration of the new test it was decided to keep the test to just one task. Messick (1992:10) points out that "under ordinary conditions of accountability assessment, trade-offs may be required between breadth of content coverage and the depth of process understanding promised by the use of extended performance tasks". The ESL staff felt that one 'longer' piece of writing could demonstrate students' writing skills more clearly than two shorter pieces, particularly under time constraints. The ELTS Validation Report (Criper and Davies, 1986:103) drew attention to the fact that giving candidates two items to complete within a limited time period may result in getting only "first draft" type essays, making differentiating between poor and good writers difficult. For it is the good writers who manage to revise and plan their essays if given a suitable time limit.

Work on the Specifications for a 'practice' writing test (Appendix 1) and the rating scheme began simultaneously by the two writing teachers to be followed by test development.

4. TEST VALIDATION PROCESS

4.1. The 'Practice' Test:

A 'practice' writing test subsequently developed (Appendix 2) was administered on the 9th and 10th of September, 1993 during the writing classes to 33 AESL students present [1]. The aims of the practice test were twofold:

(1) Gauge student and staff reactions to the new format

(2) Trial the new scoring instrument

All tests were rated by the two staff members involved in designing the test and the rating scheme. Ten randomly chosen tests were then rated by all staff.

4.2. Students' Reactions:

Students' reactions to the new format (compared to the entrance test format) were elicited orally in the classes in the week immediately following the practice test. Most students felt that the new format was preferable - both in terms of the longer time given and range of sources supplied. Students explained that the longer time enabled them to complete the test and that the range of texts allowed them to respond to the task without having to rely on any prior knowledge.

4.3. ESL Program Staff reaction:

Post-test discussions with fellow staff revealed some concerns about the nature of the excerpts chosen and illuminated the need to strike a balance between providing not enough and too much

[1] Not all students were present at the test (33 out of a total of 40).

input. There was also some concern about the test rubric; that is, it may not have set out clearly whether students needed to bring in their own personal views or whether they should just stay very closely to the input provided. These are all important considerations to be borne in mind when developing the 'real' test.

However, the new format did seem to address some of the previous concerns we had about the writing test. The new format, by giving students a number and range of sources, seemed to alleviate the over-dependency on the one text and give students the opportunity to demonstrate their synthesising skills.

4.4. Content and Face Validity

Content validity looks at "whether or not the content of the test is sufficiently *representative* and *comprehensive* [emphasis in original] for the test to be a valid measure of what it is supposed to measure" (Henning, 1987:94). In achievement-type tests, content validity is fairly straightforward as it is constrained by the goals and content of the instructional course, as set out in the specifications document.

The 'practice' test seems to have good content validity, but as mentioned in the preceding discussion, the wording of the test rubric will need more careful attention in future designs.

As to face validity, " a test is said to have face validity if it looks as if it measures what it is supposed to measure" (Hughes, 1989:27). Although, by definition, face validity is an impressionistic measure, based as it is on test takers' and test users' impressions of the test, it has been nevertheless acknowledged as having an important effect on the acceptability of the test by test takers (Bachman, 1990:288). Brindley (1986:13) notes that one of the obvious advantages of direct tests is that they are high on face validity. Gauging by the student and staff reaction described above, the new format of the writing test seems to have added to this high face validity.

5. THE NEW AESL ASSESSMENT SCHEME

5.1. The construction of the new AESL assessment scheme

It is widely acknowledged that the design of a valid and reliable scoring scheme is a very complex and arduous task. At present, the three main procedures available for directly assessing writing quality include: holistic, analytic and primary traits. The choice of procedure and of the criteria used to assess performance, the size of the scale used, the presence or absence of scalar descriptors and the wording of such descriptors, if present, are all important considerations having implications not only for the reliability of the rating judgements but also for the validity of performance assessment.

A primary trait scheme was initially contemplated as it is claimed to be more sensitive to specific context or genre features (Pollitt and Hutchinson, 1987), however, an analytic procedure was finally chosen. The choice was guided by the many apparent and relevant advantages of the analytic procedure: it can be used with multiple prompts and thus allows teachers to use it in the assessment of a wider range of tasks (Hamp-Lyons, 1991), it has proven to be the most reliable of all direct assessment procedures (Perkins, 1983; Hamp-Lyons, 1991) and above all, it can be used to provide specific diagnostic feedback to students.

The construction of the AESL rating scheme began by identifying traits considered germane to a 'good' academic essay and the language and related skills which may contribute to such a product. Other relevant rating schemes and scales were also consulted including IELTS, TEEP (Weir, 1990), a scheme developed by Brown and Bailey (1984) and one developed by Taylor and Mangelsdorf (1987) at the University of Arizona.

5.2. The AESL Rating Scheme, (version #1)

The initial version of the instrument was based on 5 categories each representing an analytic criterion:

S: Structure and cohesion;
C: Content (referring to the number and development of arguments);
V: Vocabulary and spelling;
G: Grammar and
R: Referencing skills.

Each category was in turn rated on a scale of 5 and descriptors were developed for each point. However, mid points [2] were also allowed but for these descriptors were not supplied. This, it was believed, would give raters more flexibility. Thus, the scoring scheme became a 9 point scale. The total mark (T) was then an aggregate of the 5 category scores.

The two staff members ('N' and 'J') then rated all the essays separately. The post-assessment discussion revealed some minor points which needed clarification in the descriptors, but more importantly, both raters concurred that the major weakness of the instrument was that the fifth category (R) had the potential to distort the final mark (T). That is, it could raise the mark of an essay that was otherwise rated as 'poor' on the four 'more linguistic' criteria or vice-versa [3]. It was also agreed that it would be more appropriate, given test time constraints, to assess referencing skills on assignments rather than on the test. The rating scheme was then amended accordingly.

5.3. The AESL Rating Scheme (Version # 2)

Thus on the second version of the rating scale (appendix 3) the fifth category (R) was omitted and a note added instructing raters how to treat poor referencing. All essays were then remarked by the two principal raters using the revised scheme, two weeks after the first marking session and in a different random order.

---

[2] That is, the scoring scale for each category became: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5

[3] This weakness was in fact confirmed in the language testing class (21/9/93) when a preliminary IRT analysis of the data revealed that the fifth criterion (R) was misfitting.

Furthermore, 10 scripts were randomly chosen and allocated to the other staff members for marking.

The ensuing meeting of all raters revealed that, although most raters found the scoring scheme fairly clear and easy to use, there were still some areas of concern requiring further attention. Some suggested that the distinction between a score of 3 and 4 on grammar needed to be made clearer in the descriptors. The issue of whether to mark errors in word form under grammar or vocabulary was debated and the majority felt that it should be classified as a grammatical error. The difficulty of assessing students' vocabulary was also raised, given students' tendency to rely on the language of the texts supplied. Another common concern was how many arguments should constitute a "sufficient number". These considerations will need to be incorporated in the future amended version of the rating scheme.

6. VALIDATION OF THE RATING SCHEME

6.1. Traditional Analysis

The reliability of a scoring instrument is particularly important on performance tests which use subjective ratings. Variation in scores may be due to inconsistency between raters or within raters themselves. The traditional method of assessing the reliability of the rater behaviour is to investigate inter-rater and intra-rater variability.

6.1.1. Inter-rater reliability

Inter-rater reliability measures the extent to which different raters agree about the assigned ratings. Investigations of inter-rater variability tend to focus on final scores. However, as this was a new scoring instrument, it seemed important in this study, to determine whether raters were interpreting each category on the scale consistently. Thus, inter-rater reliability was estimated for the two principal raters on the entire sample and for all 5 raters on the cohort of 10 random scripts on each rating category as well as on the aggregate score.

For all estimations of inter-rater reliability the Pearson correlation coefficient (r) was used [4]. In order to have confidence in the reliability of two raters, r is expected to be in the high 0.80s or 0.90s (Hatch and Lazaraton, 1991:441). Table 1 shows that the reliability coefficients between the two principal raters, using the first version of the scoring scheme, were unacceptably low ranging from 0.297 to 0.670 on individual categories and 0.697 for the overall score.

Table 1: Inter-rater reliability correlation matrix, Raters N & J, using Rating scheme (version #1), September, 1993 (n=33)

	N-S	N-C	N-V	N-G	N-R	N-T
J-S	0.406*					
J-C	0.204	0.297				
J-V	0.566**	0.637**	0.654**			
J-G	0.537**	0.764**	0.726**	0.670**		
J-R	0.367	0.463**	0.275	0.331	0.671**	
J-T	0.592**	0.693**	0.590**	0.627**	0.571**	0.697**

*p=<.05
**p=<.01

This could be due to the fact that this was a new scoring scheme and hence the raters were not sufficiently familiar with it. In fact, correlations estimates when the second version of the rating scheme was used (Table 2) were higher, ranging from 0.798 to 0.858 on individual categories and 0.894 on the overall score.

[4] Assumptions underlying r (i.e. normal distribution and linear relationship between the variables) were checked by analysing the scatter plots and histograms for each of the correlations and were found to be met.

Table 2: Inter-rater reliability correlation matrix, Raters N & J, using Rating scheme (version # 2) , October, 1993 (n=33)

	N - S	N - C	N - V	N - G	N - T
J - S	**0.845****				
J - C	0.566**	**0.858****			
J -V	0.622**	0.586**	**0.828****		
J -G	0.611**	0.633**	0.732**	**0.798****	
J -T	0.769**	0.832**	0.754**	0.802**	**0.894****

*p=<.05
**p=<.01

Inter-rater reliability for all raters on the cohort of 10 randomly chosen scripts revealed unacceptably low correlations [5]. Table 3 reveals that the highest correlation was on category (S) and ranged from 0.527 - 0.973; negative and very low correlations were found on all other categories.

The low correlation between rater 'C ' and all others on almost all the categories perhaps reflects the fact that this rater has been working in the ESL Program for the shortest time period. Brindley (1986:21) suggests that experienced raters often internalise a scale and base their assessment on this internalised system thus achieving higher reliability.

[5] It should be noted that these correlations statistics may not be very accurate given the small sample size. A non-parametric test such as Spearman's rho may have been more appropriate. Bachman (1990:181) recommends using coefficient alpha in such calculations; whereas Hatch and Lazaraton (1991) recommend using transformation Z and the r_{tt} coefficient to estimate the reliability of all judges rating. Another statistic which may be used is rI - a one way analysis of variance which estimates intra-group agreement rather than mere linearity.

Table 3: Inter-rater reliability correlation matrices, Raters: N, J, C, S, A, using Rating scheme (version # 2) (n=10)

3.1 Inter-rater reliability correlations on category "S" (Structure)

	N - S	J - S	C - S	S - S
J - S	0.752*			
C -S	0.527	0.769**		
S -S	0.716*	0.806**	0.636*	
A -S	0.738*	0.819**	0.686*	0.973**

3.2 Inter-rater reliability correlations on category "C" (Content)

	N - C	J - C	C- C	S - C
J - C	0.865**			
C -C	0.781**	0.926**		
S -C	0.714*	0.657*	0.700*	
A -C	0.652*	0.498	0.557	0.928**

3.3 Inter-rater reliability correlations on category "V" (Vocabulary)

	N - V	J - V	C - V	S - V
J - V	0.747*			
C -V	-0.144	0.041		
S -V	0.676*	0.447	0.062	
A -V	0.603	0.611	0.041	0.709*

3.4 Inter-rater reliability correlations on category " G" (Grammar)

	N - G	J - G	C - G	S - G
J - G	0.420			
C - G	-0.084	0.318		
S - G	0.837**	0.447	-0.181	
A - G	0.428	0.416	0.458	0.378

3.5 Inter-rater reliability correlations on "T" (Total score)

	N - T	J - T	C - T	S - T
J - T	0.744*			
C -T	0.342	0.563		
S -T	0.781**	0.609	0.450	
A -T	0.751*	0.624	0.689*	0.904**

*p=<.05
**p=<.01

6.1.2. Intra-rater reliability

Weir (1990:68) points out that inter-rater reliability in a sense assumes that raters are equally consistent in their own individual assessment over time. However, in any rating situation, certain factors such as rating sequence may affect the rater's consistency over time.

Intra-rater correlations (using Pearson's r) were estimated for each principal rater and on each category using version 1 and 2 of the scoring scheme [6] . Table 4 shows that intra-rater reliability coefficients for rater 'J' tended to be moderate to good but still significant; Table 5 shows that they were significant and good for rater 'N' . Thus it seems that both raters were fairly consistent in their ratings.

Table 4 : Intra-rater correlation matrix, Rater: J (n=33)

	S_1	C_1	V_1	G_1	T_1
S_2	**0.758****				
C_2	0.453**	**0.629****			
V_2	0.329	0.288	**0.786****		
G_2	0.210	0.283	0.773**	**0.862****	
T_2	0.502**	0.507**	0.789**	0.851**	**0.845****

[6] This was possible despite the fact that 2 versions of the scheme were compared since apart from omitting (R) as a criterion, very little changes were made to version # 2 of the scheme.

Table 5: Intra-rater correlation matrix, Rater: N (n=33)

	S_1	C_1	V_1	G_1	T_1
S_2	**0.832****				
C_2	0.741**	**0.803****			
V_2	0.700**	0.677**	**0.835****		
G_2	0.722**	0.766**	0.862**	**0.884****	
T_2	0.849**	0.816**	0.844**	0.847**	**0.922****

*p=<.05
**p=<.01

6.1.3. Discussion

Rater training and the practice of double marking are means of improving the reliability of subjective ratings. The above findings show that both are needed in the ESL Program if the new rating scheme is to be used in future assessment decisions.

However, correlation estimates should be interpreted cautiously. Mullen's (1980) study has shown that reliability and differences in scores can co-exist; that is, scores can be closely parallel but not equivalent. Furthermore, inter-rater reliability estimates do not indicate whether different raters (or for that matter the same rater over time) are evaluating each criterion independently or whether they are being influenced by some criteria more when making assessment decisions.

6.2. Rasch IRT Analysis

6.2.1. Introduction

The previous analysis allows us to make statements about the scoring scheme's reliability; however, reliability is "a necessary but not sufficient condition for validity to be present" (Henning, 1987:89). The Rasch IRT Model represents a relatively new, probabilistic approach to the analysis of test data which provides information about the properties of the test, test takers and test items.

The Rasch IRT Model has many advantages over classical analysis of test data. One advantage is that it links item difficulty and person ability and maps them onto the same probability scale, measured in logits, interval level units. This visual representation makes it easier to draw conclusions about the difficulty or otherwise of a particular test for a particular group of candidates. The second and perhaps the greatest strength of the Model is its ability to generate inferences from the data in the form of estimates of candidates' underlying ability independent of the particular test items, and of the underlying difficulty of test items independent of the abilities of a particular trial group (McNamara, in preparation). This latter aspect of the analysis is particularly relevant in the validation of a scoring scheme, for in the analysis of a writing test the rating categories can be treated as test items.

Thus in this study, it is the scoring scheme which will be the main focus of investigation, in particular looking at its ability to calibrate students' ability appropriately, the nature of the categories and the scales used and most importantly, its construct validity; that is, can the four individual scores be added up to yield a valid overall score of writing proficiency? These are all important issues to consider if the scheme is to be used in the future assessment of all writing tasks in the AESL course. Discussion will be with reference to measure estimates and to reported fit values for both items and persons (cases).

6.2.2. Data Used

The data used for the analysis was the scores allocated to the 33 candidates on each rating category using the second version of the assessment scheme [7].

[7] Raters expectations of their students have been shown to have an effect on their assessment (Huot, 1990:255). As both raters ('N' and 'J') are also the writing teachers, it was decided to use the rating given by each rater to students for whom the rater is not the writing teacher. Thus, for candidates 01-16 rater J's scores were used and for candidates 17-33 rater N's scores were used.

Table 6, sets out the frequency of scores given on each analytic criterion (rating category) and shows the presence of the anticipated 'shrinkage factor' (Carroll and Hall, 1985:78); that is, the lowest points on the scale (1, 1.5) and the highest (5) were rarely used.

Table 6: Response frequencies
October, 1993
N=33

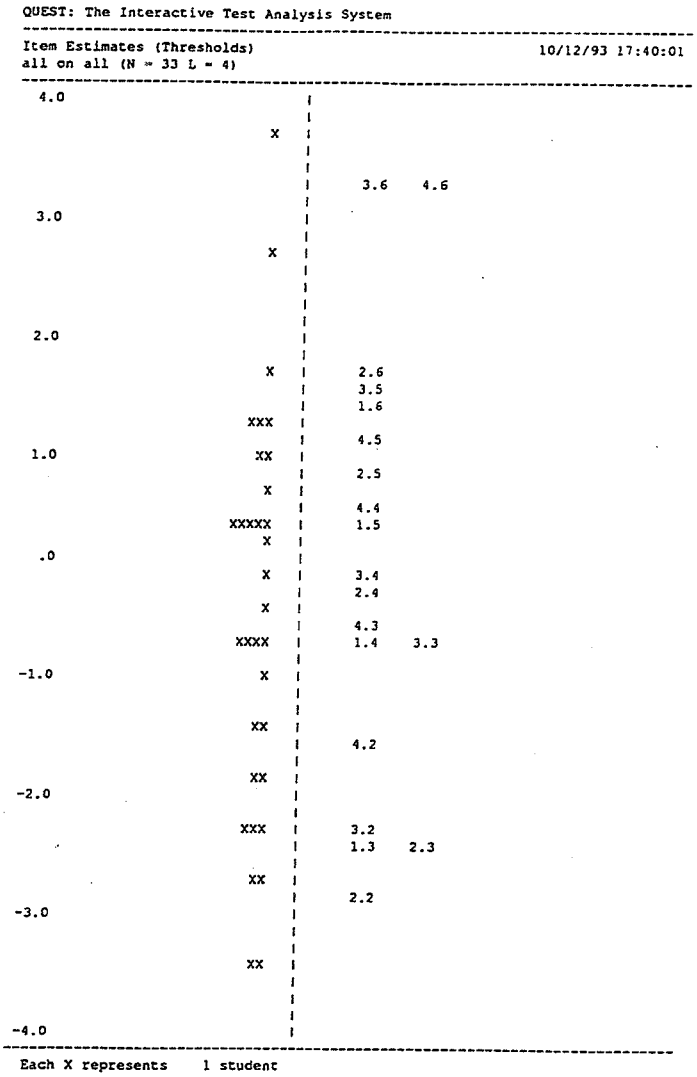
score categ.	1	1.5	2	2.5	3	3.5	4	4.5	5
S	0	0	0	5	9	7	6	6	0
C	0	0	3	1	13	7	5	4	0
V	0	0	6	8	4	10	4	1	0
G	1	0	8	6	8	3	6	1	0

Thus the scale was collapsed to a 6 point scale, the lowest 3 scores being incorporated into a single score. The data was then recoded as 1,2,3,4,5,6 to represent the range of scores used (from 2 to 4.5).

6.2.3. Information about the test

Figure 1 (Item Estimates Thresholds) is an item-ability map: a graphical representation of item difficulty and item ability mapped onto the same logit scale. It demonstrates that the AESL writing test achieved a good coverage of the ability range. Approximately half of the candidates are above and half below the average item difficulty (set at 0 logits). It also shows that the test was challenging for most candidates and at the appropriate level of difficulty. There was only one candidate who was above the level of difficulty of the test (whose ability was greater than 3.72 logits).

Figure 1


QUEST: The Interactive Test Analysis System
Item Estimates (Thresholds)
all on all (N = 33 L = 4)
10/12/93 17:40:01
4.0
3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
3.6 4.6
2.6
3.5
1.6
4.5
2.5
4.4
1.5
3.4
2.4
4.3
1.4 3.3
4.2
3.2
1.3 2.3
2.2
Each X represents 1 student

6.2.4. Information about the candidates: person measures

The reliability estimate of person measures, that is "the proportion of the observed variance in measurement of ability which is not due to measurement error" (McNamara, 1990b: 56) is acceptably high at 0.85.

Table 7 sets out the estimates of person measures (case estimates) ranging on the logit scale from -3.41 to 3.72. However, since this is a very small sample [8] the errors associated with these measures of person ability are very high. Thus inferences of underlying measures of candidates' abilities on the basis of this limited data need to be very cautious.

Table 7

QUEST: The Interactive Test Analysis System

Case Estimates In Input Order 10/12/93 17:40:14
all on all (N = 33 L = 4)

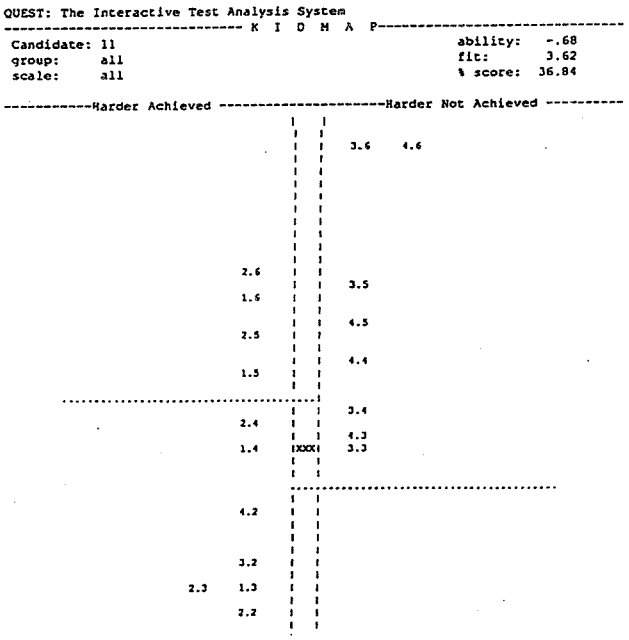
NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t
1 01	14	19	1.27	.57	1.49	1.30	.86
2 02	12	19	.68	.52	.93	1.04	.08
3 03	5	19	-1.36	.62	1.09	1.36	.38
4 04	1	19	-3.41	.97	.84	1.10	.12
5 05	11	19	.42	.51	.22	.20	-1.58
6 06	14	19	1.27	.57	.63	.54	-.42
7 07	11	19	.42	.51	2.41	2.38	1.77
8 08	9	19	-.11	.52	.30	.34	-1.19
9 09	2	19	-2.72	.75	.73	.57	-.41
10 10	8	19	-.38	.53	.39	.44	-.95
11 11	7	19	-.68	.55	3.62	4.11	2.56
12 12	3	19	-2.21	.69	.53	.58	-.50
13 13	17	19	2.73	.88	.10	.12	-1.36
14 14	10	19	.16	.51	.44	.48	-.82
15 15	13	19	.96	.54	1.82	1.97	1.29
16 16	4	19	-1.77	.65	.19	.19	-1.18
17 17	3	19	-2.21	.69	.40	.43	-.78
18 18	2	19	-2.72	.75	1.13	.96	.42
19 19	14	19	1.27	.57	.65	.61	-.39
20 20	18	19	3.72	1.15	1.95	4.12	1.19
21 21	5	19	-1.36	.62	.99	.89	.26
22 22	7	19	-.68	.55	.96	1.51	.17
23 23	4	19	-1.77	.65	.63	.64	-.20
24 24	7	19	-.68	.55	1.05	1.06	.30
25 25	1	19	-3.41	.97	.72	.78	-.03
26 26	11	19	.42	.51	1.33	1.16	.67
27 27	3	19	-2.21	.69	.40	.43	-.78
28 28	6	19	-1.00	.58	.52	.47	-.56
29 29	11	19	.42	.51	.13	.12	-2.03
30 30	13	19	.96	.54	.40	.32	-1.10
31 31	15	19	1.63	.63	.98	.86	.23
32 32	11	19	.42	.51	1.22	1.20	.53
33 33	7	19	-.68	.55	.19	.22	-1.63
Mean			-.38		.89	.99	-.15
SD			1.70		.73	.96	1.02

[8] Henning (1987:116) claims that the Rasch Model is fully operative with a minimum sample size of 100 persons.

More importantly to this study is the consideration of fit values as a means of investigating the scoring scheme's credibility; that is, how well it measures the ability of all candidates. Fit values (denoted by the t statistic) larger than +2 indicate a significant deviation from the expected response pattern and are referred to as 'misfitting' person measurements, that is, a measurement lacking consistency or as Henning (1987:123) defines it "lacking response validity". Only one misfitting person measurement is found in the AESL writing test data - candidate # 11 [9].

The 'kidmap' produced for candidate #11 (figure 2) shows in the upper left quadrant items which were estimated to be too difficult for the candidate but which the candidate attained, hence the high t value.

Figure 2



[9] Although only one candidate was identified using t values, this does represent almost 3% of the sample which is beyond the level of acceptability recommended by Pollitt and Hutchinson (1987:82). Furthermore, if we use the range of INFIT MNSQ ±2 SD as a guide 2 candidates can be identified as having misfitting person measurement.

The raw scores for this candidate (S:4; C:4; V:2; G:2) suggest that the large difference of 2 points between the scores on item 3 and 4 (V and G) compared to category 1 and 2 (S and C) contributed to this discrepancy. No other candidates' scores demonstrate such a large gap between category scores. Thus the actual scores identify an ESL learner who has acquired different components of written control at different rates - a not unusual phenomenon in ESL classes (Hamp-Lyons, 1991:241).

However, in terms of measurement, what the misfitting statistic may imply is that the raters are assessing the candidate's performance on some categories in relative terms: assessing a particular dimension as either very strong or very weak in relation to performance on other categories "rather than relative to the performance of other candidates in the categories concerned, as the scoring procedure requires. The score pattern was thus identified by the analysis as improbable" (McNamara, 1990b: 58). In other words, the scoring scheme may not be assessing this candidate's underlying ability appropriately.

6.2.5. Information about the measurement scheme: categories and scales

In the analysis of a writing test, as mentioned previously, the rating categories are treated as items and thus in the estimation of item difficulty both category and scale are taken into consideration. The logit scale used by the Rasch Model, being an interval scale, allows for valid comparisons between categories in terms of their difficulty measure.

Figure 1 clearly shows that items 3 and 4, representing the criteria of vocabulary and grammar respectively, were the most difficult items. They are consistently higher on the logit scale than the other two categories at each score level [10]. They were also the hardest categories on which to attain the top score (6) [11].

---

[10] For example, if we compare 4.6 ; 3.6 to 2.6; 1.6 or 4.3; 3.3 to 2.3; 1.3

[11] The mean ability estimate needed to attain a 6 on category 3 or 4 is 3.72 as compared to 1.56 on category 2 or 1.84 on item 1.

The fact that these are the hardest categories in which to attain a 'perfect' score may be related to the generally recognised reluctance of raters to award such scores which are normally equated with 'native speaker proficiency' to ESL learners.

The finding may also be explained in terms of the test itself . In this new format the input for the content was in fact provided. Furthermore, as an achievement test it reflects the instructional program where the focus was on academic essay writing skills more than on the development of linguistic skills. Pollitt and Hutchinson (1987:86) found in their study that expression (syntax and lexis) were hardest marked compared to appropriacy (choice of style, conventions) and offer a similar explanation.

Table 8, Item Estimates (Thresholds) suggests that (despite the high errors associated with these estimates) the scale may not be equidistant. This in turn would have some implications for the validity of adding the scores in producing a total score representing the candidate's performance. Davies (1992:13) in fact states that in rating instruments "equal interval scales is a myth."

Table 8

Item Estimates (Thresholds) In Input Order 10/12/93 17:40:09
all on all (N = 33 L = 4)

	ITEM NAME	SCORE	MAXSCR	THRESHOLD/S 1	2	3	4	5	6	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	Item 1	65	132			-2.44	-.68	.32	1.34	1.19	1.08	.8	.4
						.88	.80	.79	.82				
2	Item 2	87	165		-2.78	-2.44	-.28	.05	1.74	1.05	1.51	.3	1.5
					.98	.97	.77	.84	.88				
3	Item 3	67	165		-2.19	-.70	-.08	1.50	3.30	.69	.74	-1.2	-1.0
					.84	.81	.82	.94	1.44				
4	Item 4	60	165		-1.56	-.52	.55	1.07	3.35	.72	.60	-1.1	-1.5
					.78	.80	.80	.83	1.46				
Mean				0.00						.91	.99	-.3	-.1
SD				.56						.25	.41	1.0	1.4

6.2.6. Construct Validity

The rating instrument is a construct which is "encoded in the wording of the rating scale" and the criteria used to assess performance, "constitute an implicit view of language proficiency" (McNamara, in preparation: 28) The construct of language proficiency implicit in this scheme is that the measure of language proficiency is an aggregate of the identified component criteria, where each criterion contributes independently to the overall proficiency measure. It is this belief which allows raters to add up scores on the four categories and report a candidate's proficiency as a single score.

Item fit statistics investigate the validity of this construct; questioning whether it makes sense to add scores from different criteria ratings. The presence of any misfitting items would indicate that such scores cannot be added as such items are not measuring the same underlying ability.

The graphical representation of item fit statistics in this study (figure 3) show no misfitting items (i.e. MNSQ INFIT > 1.3) thus demonstrating that this is a valid construct of writing proficiency; that is, that the four scores can be justifiably aggregated to yield a score (T) and which in turn represents a valid measure of a candidate's academic writing proficiency.

Figure 3

```
QUEST: The Interactive Test Analysis System
-------------------------------------------------------------------------------
Item Fit                                                    10/12/93 17:40:06
all on all (N = 33 L = 4)
-------------------------------------------------------------------------------
INFIT
 MNSQ       .63       .71       .83      1.00      1.20      1.40      1.60
-------------+---------+---------+---------+---------+---------+---------+----
  1 item 1                    .               |        *      .
  2 item 2                    .               |  *            .
  3 item 3           *        .               |               .
  4 item 4                *   .               |               .
===============================================================================
```

The presence of any overfitting items (MNSQ INFIT <0.75), on the other hand, indicate a "lack of independence between scores for such an item and scores on other categories." (McNamara, 1990a:390). Figure 3 shows items 3 and 4, representing the criteria of vocabulary and grammar respectively, as clearly 'overfitting'.

That both are overfitting is perhaps not surprising given that the raters admitted to having difficulties in distinguishing clearly between a grammatical and a lexical error and hence in rating these errors on the appropriate category. Some schemes in fact do include both syntax and lexis under the same category (eg. Pollitt and Hutchinson, 1987:75). The new revised IELTS scheme has also collapsed vocabulary and grammar into one category.

What these overfitting items seem to indicate is that the raters' overall assessment of writing proficiency may be influenced by their perceptions of the candidate's grammatical and lexical accuracy [12].

Brown and Bailey (1984) review a number of studies which have attempted to determine the relationship between a grammatical measure and overall scores and conclude that the research findings are contradictory. Homburg (1984) in fact argues that measures of sentential grammar may be more influential at the lower levels of writing proficiency but that discourse measures become more influential as proficiency increases. Mullen's study (1980) on the assessment of University students' ESL essays using an analytical score found that it was ratings on vocabulary usage which accounted for 84% of the variance in the overall score. In his investigation of the Occupational English Test (OET), McNamara (1990a:65) found that grammar and appropriateness contributed to 66% of the variance but that grammar itself accounted for 60% of the variance.

McNamara (1990a: 397) also claims that raters' orientation to grammatical accuracy is "very deep-seated" being the result of the

---

[12] A stepwise regression analysis or part-to-whole correlations would need to be carried out in order to verify this conclusion.

training of raters as language teachers and hence tends to be impervious to rater retraining.

7. CONCLUSION

There is no doubt that the criteria used to assess essays are interdependent and that it is inevitable that raters will be influenced in their judgement of any one criterion by the qualities of others. However, the evidence that it is one (or two) particular categories which drive the assessment of all other categories consistently is perhaps of grave concern for it has implications for both the assessment and the course of instruction.

In terms of assessment, the adding up of the scores from the four categories to yield a total score is based on the assumption of equal weighting. The analysis has shown that vocabulary and grammar may, in fact, be inadvertently receiving greater weighting in this rating procedure.

Huot's review (1990) concluded that the majority of research indicates that raters are mostly concerned with content considerations when rating compositions, but admits that whether this belief concurs with how they actually score the paper needs further investigation.

In this study, the list of enabling skills listed in the Specifications document (Appendix 1) and the deliberations preceding the construction stage of the new rating scheme reflected a concern with content and structure considerations. Yet, what this analysis seems to have shown is that raters are mainly influenced by lexical and linguistic accuracy and appropriacy in their rating, without even being aware of the bias in their assessment behaviour.

The findings also have pedagogical implications particularly if the test professes to be an achievement test. That is, if it is the candidates' control of lexical and grammatical features which determines their overall score more so than any other feature of their writing and that these two categories have been shown to be

the hardest categories, then perhaps more time needs to be devoted in classes to improving these skills.

The rating scheme developed obviously needs further refinement and trialing prior to being used on a larger scale. More importantly, its use must be preceded by a session in which the findings of this study, not only in terms of rating reliability, but more in terms of the scheme's validity and its implications for the goals, standards and focus of the AESL course are discussed openly.

REFERENCES

Bachman, L. F. (1990) Fundamental considerations in language testing. Oxford: Oxford University Press.

Brindley, G. (1986) The assessment of second language proficiency issues and approaches. Adelaide, SA: National Curriculum Resource Centre, Adult Migrant Education Program Australia.

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

Carroll, B. J. (1980) Testing communicative performance. Oxford: Pergamon.

Carroll, B. J. and P. J. Hall (1985) Make your own language tests: A practical guide to writing language performance tests. Oxford: Pergamon.

Criper, C. and A. Davies (1986) Edinburgh ELTS validation project. Project report. Edinburgh: Department of Applied Linguistics, University of Edinburgh.

Cumming, A. (1990) Expertise in evaluating second language compositions. Language Testing 7, 1: 31-51.

Davies, A. (1992) Is language proficiency always achievement? Melbourne Papers in Language Testing 1,1: 1-11.

Greenberg, K. (1986) The development and validation of the TOEFL Writing Test: A discussion of TOEFL Research reports 15 and 19. TESOL Quarterly 20, 3: 531-541.

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

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.

Johns, A. M. (1986) Coherence and academic writing: some definitions and suggestions for teaching. TESOL Quarterly 20,2: 247-265.

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

McNamara, T. F. (1990b) Item Response Theory and the validation of an ESP test for health professionals. Language Testing 7,1: 52-75.

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

Messick, K. A. (1992) The interplay of evidence and consequences in the validation of performance assessments. Conference paper given at the Annual Meeting of the National Council on Measurement in Education, San Fransisco, April, 1992.

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, 160 - 170.

Perkins, K. (1983) On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. TESOL Quarterly 17, 4: 651-671.

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

Shih, M. (1986) Content-based approaches to teaching academic writing. TESOL Quarterly 20, 4: 617-648.

Taylor, V. and K. Mangelsdorf (ed) (1987) The students guide to ESL composition. Arizona: Composition Program, English Department, University of Arizona.

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

APPENDIX 1

AESL WRITING TEST SPECIFICATIONS
(practice test)
September, 1993

1. General statement of purpose:

An achievement and proficiency test: testing academic writing proficiency directly via a test which requires student to write an academic argumentative essay thereby demonstrating a mastery of the skills covered in the AESL course work.

Skills tested:

(i) Ability to organise ideas into an academic essay format: with a clear introduction, body and conclusion.

(ii) Ability to develop an extended argument.

(iii) Ability to extract and synthesise[1] information from a given range of academic and non-academic texts (stimulus).

(iv) Ability to acknowledge sources used appropriately.

(v) Ability to write coherently and cohesively

(vi) Ability to express oneself accurately and appropriately using a range of sentence structures and vocabulary.

2. Target population

Students from the Advanced English as a Second Language Course, about to complete a 1 year content-based EAP course at Melbourne University. Students' writing proficiency is approximately in the range of 5 - 7 on IELTS writing bandscales.

---

[1] Synthesis of information refers here to the combining or contrasting of information from secondary sources and incorporating it into the writer's scheme.

3. Format: Texts + 1 essay question[2]

(i) a set of short reading extracts will be supplied presenting a range of views/issues relating to the given topic. Students may use any of the evidence presented in their essay and/or add to it.

(ii) One essay question on an argumentative topic[3]

4. Test sections:

(i) Stimulus: a set of brief articles/extracts
(ii) 1 Test question
(iii) Written Instructions to candidates outlining the requirements of an academic essay, stating the required length and recommending times for each activity.

5. Stimulus material

A range of texts/extracts drawn from authentic academic texts and/or media reports[4] presenting a range of ideas or perspectives on an argumentative issue. The Texts drawn from academic sources should come from authentic university level texts. Any overtly subject-specific or technical vocabulary should be glossed in the footnotes. Texts from the media should also aim at a university educated audience.

Texts length: some texts could be as short as a 1 sentence quote but the majority should be at least one paragraph long but preferably no longer than 2 paragraphs (per text). The total length of stimulus material should not exceed 1.5 pages.

---

[2] As this was a practice test only one essay topic was given. It was felt that this would make the trialing of the marking instrument more straightforward by eliminating possible variability due to topic choice (Mullen, 1980).

[3] In the Second semester students worked on argumentative essays in the writing classes.

[4] The number of texts drawn from each source may depend on the topic chosen and the availability of suitable texts. It is recognised, however, that

6. Item type

1 essay question on an argumentative issue which requires students to consider arguments for and against the given issue.

The issue chosen should relate to the issues/content covered in the second semester of the AESL course[5] , but should not necessarily presume prior extensive knowledge about the topic.

7. Response Attributes

Candidates are asked to write approximately 2 pages (about 500 - 750 words) on the given topic. Answers may be based purely on the input provided (texts) but students are invited to express their own views on the issue.

8. Response time

Total time allocated for test: 90 minutes

Recommended time per activity (as specified in the instructions to candidates):

Reading:	15 - 20	minutes
Planning:	10 - 15	minutes
Writing:	45	minutes
Editing:	10 - 15	minutes

---

the predominance of one source or genre could alter the complexity of the test.

[5] The topics covered in the second semester include:

(i) Issues in the constitutional debate
(ii) Sources of Australian Law and legal issues around us
(iii) The Australian system of government , political parties & ideologies
(iv) Multiculturalism and the Law
(v) Australian International Relations: historical perspectives & contemporary concerns

9. Marking

All essays are to be double marked using the new analytic scoring instrument (currently developed by ESL Program staff). Each criterion is to be marked on a 9 point scale (1 - 5; but allowing for mid-points as well).

The 5 analytic criteria used to assess essays are:

1. S: Structure and cohesion
2. C: The number of ideas and their development
3. V: Appropriacy and range of vocabulary and correct spelling
4. G: Grammatical accuracy and range of sentence structures
5. R: Use and appropriate acknowledgment of sources used

10. Reporting

Students will receive a diagnostic profile report on their essays, highlighting their strengths and weaknesses under each analytic criterion.

APPENDIX 2

AESL, September, 1993

Timed Writing

Using information from the extracts provided and your own ideas, write an academic essay on the following topic:

In Australia, there are progressively more and more bans on smoking in public places; yet smoking is still legal. What are the economic considerations in totally banning the growing and selling of tobacco products in Australia and/or worldwide?

Time: 1 1/2 hours

Length: approx. 2 pages (500 - 750 words)

Instructions:

1. Spend approximately 15 - 20 minutes reading and 10 minutes planning for your essay. Plans should be submitted with your essay.

2. Your should write for about 45 - 60 minutes.

3. Any ideas taken from the reading extracts should be appropriately acknowledged.

4. Spend 10 - 15 minutes re-reading and editing your essay

Tobacco is easy to grow and provides a ready source of cash to the small farmers who still constitute most of the world's producers.

United Nations Food and Agriculture Organisation, Yearbook 1977, Volume 31. Rome, UNFAO, 1978, p.290.

Tobacco becomes a net cost to society whenever a large proportion of the population smokes enough to suffer the impact of tobacco-induced diseases.

Tobacco - Hazards to Health and Life. NSW Cancer Council, Position Paper, 1985, p.21.

The land used to raise tobacco is not available to raise food, and this too may contribute to malnutrition and higher mortality in developing countries.

United Nations Food and Agriculture Organisation, Yearbook 1977, Volume 31. Rome, UNFAO, 1978, p.38.

In China, about 60% of the price of a packet of cigarettes goes to the government in the form of taxes.

Mathews, J. Between Puffs, Chinese are told of Cigarette Perils. Washington Post, 6/9/78, p.17.

Thousands of small retail traders would be severely limited in their ability to continue their business and to employ shop assistants if not for sales of tobacco.

Small Retailers Association Report, 1982, p.42.

In the United States at least, where tobacco was first developed as a colonial product for export, the health costs of domestically consumed tobacco now far outweigh the dollar returns to producers, manufacturers, exporters, and tax collectors. While total consumer spending (plus exports) now amounts to about $19 billion and supports jobs for 1.3 million people, the cost to US citizens in lost production from sickness, health care, and loss of life and property destroyed by fire totals $27.5 billion.

Miller, R.H. The Economic Importance of the US Tobacco Industry. Washington D.C., US Department of Agriculture, May 1982, p.187.

...it is sometimes argued that the government would save money in the absence of smoking since it would not have to pay certain sickness benefits to smokers and pensions to spouses of deceased smokers. Against this are old age pension savings which occur if smokers' life expectancy is lower than that of non-smokers.

Hunt, B. (1987) Submission to the Industries Assistance Commission Investigation of Tobacco Growing and Manufacturing Industries. Appendix 3: Measures affecting tobacco consumption, Tobacco Institute of Australia, page 12.

In November 1985 the Prices Surveillance Authority reported that the three major tobacco manufacturers operating in Australia achieved, on average, higher profits than Australian industry generally, and had been one of the more profitable industries for many years. This was attributed to the industry's relative immunity to the economic recession, largely because of the comparatively inelastic nature of demand for cigarettes, and advancements in mechanisation.

Prices Surveillance Authority. Report No. 6 - Inquiry in relation to the supply of cigarettes, Matter no PI/85/2, 26 November 1985

APPENDIX 3

Scoring Scheme for AESL writing

(NOTE: Whole points or mid points may be allocated)

S: Structure & cohesion*	C: Content	V: Vocab. & spelling	G: Grammar
5: Well structured and organised essay; thesis clearly stated in introduction, good use of transitional expressions; logical arrangement of content; conclusion complete.	5: Good range of ideas/arguments addressing the topic assigned. The ideas are concrete and thoroughly developed and evaluated. A sophisticated , original synthesis of issues.	5: Excellent range of vocabulary, used appropriately and accurately. Good register.	5: Excellent range of sentence structures .All structures accurate and appropriate.
4: A well defined structure. A good introduction clearly stating thesis/plan and . The conclusion is complete and appropriate to the essay. There may be some inconsistencies in use of transitions or in paragraphing.	4: Overall a good range of arguments presented and supported. OR a limited number of arguments /ideas but very well developed/supported There is a synthesis and an evaluation of arguments/ideas. There may be some minor inconsistencies in the development or relevance of some arguments.	4: Good range of vocab; appropriately used. Inconsistent errors in usage or minor spelling errors may still be present.	4: Overall good range of sentence structures; some inconsistent errors may be present.
3: A satisfactory structure but may be uneven in terms of coherence and/or cohesion. The introduction has a clear thesis or plan announced and an acceptable conclusion. Some flaws in paragraphing or in cohesion within paragraphs.	3: A sufficient number of arguments/ideas considered, but may still be either poorly developed/ supported or irrelevant. There is an attempt at synthesis and /or evaluating arguments but this attempt is of an uneven quality.	3: Range of vocab fairly good. Inadequacies may be related to verbosity / appropriacy/ minor spelling errors.	3: Satisfactory range of sentence structures. However, some structures may be inappropriate. Grammatical errors : either a few instances of a wide range of error types or consistent (i.e. systematic) errors from a narrow range of error types evident.
2: There is an obvious attempt at structuring, but structure is poor. Poor coherence or transitions absent or inappropriate. Severe flaws in introduction i.e no clear thesis statement or plan. A minimal or an inappropriate conclusion. Poor paragraph structure.	2: Very few arguments presented. Those presented are either treated fairly superficially (e.g. noticeable over-reliance on quotes). Arguments/ideas may not be well developed or well supported (e.g misplaced quotes). Some arguments not relevant to topic. There may be an attempted synthesis, but attempt is poor.	2: Range of vocab. limited or often inappropriate. Spelling obviously a problem.	2: Attempts at variety of sentence structure may be evident but often with inaccuracies and punctuation errors. An unacceptable frequency of surface level errors such as word forms, agreement, articles, tense use and formation.
1: No apparent organisation of content . Absence of introduction.	1: Not many ideas; those included are either irrelevant, insufficiently developed or not supported at all. Almost total plagiarism.	1: Poor word choice or severe spelling problems. Or work largely plagiarised hence difficult to assess candidate's vocab.	1: Limited range of sentence structures and/or very frequent grammatical errors often impeding comprehension. Unintelligible sentence structure.

* For assignments: Deduct 0.5 of a mark if bibliography or referencing is poor or 1 if absent