Inter-rater Reliability in an ESP Context

Sheryl Ward

Abstract

This paper discusses the issue of inter-rater reliability in pre- and post-tests in an ESP context: a 3-day workshop on technical report writing for senior Chinese engineers in Hong Kong. In particular, the paper addresses two questions of inter-rater reliability. First, to what extent experienced raters' assessments are reliable when based on a subjective, qualitative approach rather than an explicit, quantitative approach, and second, to what extent language experts' assessments are similar to a subject expert's assessment of the same writing tasks. The two experienced trainers who conducted the workshop rated the twelve pre and post-tests using a largely qualitative approach. Three months after the workshop, a senior manager of the company also rated the twelve post-tests using a simple rating scale that he developed specifically for the assessment task. One of the trainers then rated the post tests again using this rating scale. The results of this case study indicate that raters from both linguistic and non-linguistic backgrounds focus on similar criteria when assessing second language writing samples. However, even though there is agreement about the salient criteria to be used for assessment purposes, inter-rater reliability, first, between the two language experts and second, between the language and subject experts, is not strong. It is argued that the use of an explicit rating scale for the trainers and some rater training for the manager would have improved the inter-rater reliability for both groups of raters. It is concluded that the assessment phase of the training cycle could benefit from the contributions of both language and subject experts.

1. Introduction

In a results-oriented business world, it is increasingly necessary to show that money spent on skills training has been put to good use. Therefore, management often requires some measure of achievement from training courses. One method of measuring achievement in such courses is to conduct pre- and post-tests, whereby participants are tested prior to the course to establish base-line data and then tested

Melbourne Papers in Language Testing 1997 Volume 6.2 pp. 73-91. The Language Testing Research Centre, The University of Melbourne.

again at the completion of the course. This paper explores the issue of inter-rater reliability in pre- and post-tests in an ESP context: a 3-day workshop on technical report writing for senior Chinese engineers in Hong Kong. In particular, the paper addresses two questions of inter-rater reliability: first, to what extent experienced raters' assessments are reliable when based on a subjective, qualitative approach rather than an explicit, quantitative approach, and second, to what extent language and subject experts' assessments are similar.

2. Reliability

Reliability is the extent to which the results of a test can be produced consistently.

> *This kind of accuracy is reflected in the obtaining of similar results when measurement is repeated on different occasions or with different instruments or by different persons (Henning 1987:73).*

One way of establishing the reliability of productive tasks - speaking and writing - is to measure inter-rater reliability. To establish reliability among raters, it is usual practice to decide on explicit assessment criteria, to train raters and then to conduct periodic monitoring in order to ensure consistency in ratings. Ideally, a correlation level of at least 0.8 is necessary to be confident that raters are assessing the same task in the same way.

However, in a business-training context, it is usual to have only one rater (usually the trainer) assess end-of-course productive tasks. Rater training and ongoing monitoring are therefore seldom considered practical or relevant options. Explicit rating schedules with clear criteria would seem to be even more essential in this context to avoid inconsistent judgements by a sole rater. As Henning points out:

> *Any rater called upon to make subjective estimates of composition quality or speaking ability in a language is liable to be inconsistent in judgement. This is particularly true in situations where the raters are not provided with detailed rating schedules (Henning 1987:76).*

This paper explores first what happens when two experienced linguistic experts independently assess the same writing samples without using a detailed rating schedule. It then explores what happens when a 'linguistically naive' rater assesses the same writing samples using his own rating schedule.

3. The Study Context

3.1 The Workshop

The author was asked to design and deliver a 3-day report writing skills workshop for senior engineers who worked for a large transportation consultancy in Hong Kong. Management had been receiving complaints from clients about the poor quality of the reports submitted by these engineers.

3.2 The Participants

The twelve participants in the workshop were all native Cantonese speakers from Hong Kong. All were graduates of Hong Kong or overseas universities. All were reasonably fluent English speakers but had some difficulties in writing English. Writing reports was a large part of the job description for each of these participants.

3.3 The Raters

Two experienced language trainers co-taught this workshop and conducted the initial assessments of the pre- and post-tests. The two trainers had worked closely together for two years prior to the workshop as English language trainers for a large international bank in Hong Kong. They were thus very familiar with the target group and the cultural context.

In addition, a senior director of the company also assessed the tests. This director is referred to as the 'subject expert' in this article, while the two trainers are referred to as the 'language experts'. The subject expert is a native English speaker and is regarded within the company as being one of the 'best' report writers. He is an engineer, has had no teaching background, and was given no training prior to assessing these tests. Neither was he provided with criteria on which to base his assessments. Instead. he was asked to establish his own assessment framework. The main purpose of the second part of the study was to compare the approach, criteria and assessments of the subject expert with those of an experienced language rater. Research on rater behaviour often suggests that linguistically naive native speakers vary considerably, and unpredictably, in their perceptions of foreigner talk with respect to the dimensions along which they evaluate performance and the degree of consistency and tolerance they manifest in their judgements (Elder 1992:16)

4. The Pre and Post Test

The test used was adapted from an example of a memo report in Huckin and Olsen (1991: 238). The participants in the workshop were asked to write the Introduction and Management Summary for this report. They were asked to complete this task at the beginning of the three-day course and again at the end of the course. The pre-tests were collected by the trainers and then evaluated along with the post tests.

5. Assessment Procedures

The pre- and post-tests were not 'high stakes' tests in that their main purposes were to indicate to management some measure of improvement in the participants and to offer encouragement to the participants about their progress in writing. As such, a more informal, less systematic approach to assessing the tests by the raters is justifiable.

Prior to assessing the pre- and post-tests, the two trainers informally discussed and agreed on the criteria on which they would base their assessments. These criteria were based on the three themes of the workshop:

Clear Focus: The writer can state the purpose of the report clearly and explicitly; information is organized in a logical manner .

Clear Structure: The writer has included all information relevant to these sections of the report (ie Introduction and Summary).

Clear English: The writer has expressed him/herself in sentences that are coherent, grammatically correct and clear.

Each trainer then independently assessed seven (ie 7 pre and 7 post tests) of the twelve tests by writing brief comments about salient features of each of the tests. One trainer then collated the two sets of comments and wrote a short report on each of the seven participants. A brief discussion was held between the trainers prior to writing these reports to discuss differences between their assessments of one of the tests. One trainer then assessed the remaining five pre- and post-tests and wrote short reports on them as well.

6. Limitations

Three limitations of this approach are immediately obvious:

- The criteria are too general and are not mutually exclusive. For example, comments related to the purpose of the report could be included under any one of the three criteria.

- An explicit, detailed scoring schedule was not used - holistic comments were generally used instead.

- Inter-rater reliability cannot be established systematically - the closest the raters came to establishing inter-rater reliability was the collation of the two sets of comments and a discussion regarding the results of one of the tests.

Nevertheless, statistical analysis using a Chi square test indicates that the two raters were focusing on similar criteria during their independent evaluation of the tests. In addition, where the comments between the raters matched, they agreed more than they disagreed.

7. Issue 1: Inter-rater reliability between language experts

This section of the paper explores two questions:

1. To what extent do the trainers focus on the same criteria without using an explicit, detailed rating schedule?

2. To what extent do the trainers agree about the quality of the writing, based on the comments they made during their assessment?

8. Analysis

A post-hoc analysis of the two raters' comments was made. These comments were classified into eight categories (Table 1). These eight categories were later collapsed into four so that a Chi-square test could be carried out. The results of this test are shown in Table 2. To determine the extent of agreement between the ratings, each trainer's comments on each participant and each test were analysed in terms of whether they agreed, disagreed or did not match at all. These results are shown in Table 3.

Agree = Both Rater 1 and Rater 2 comment on the same criteria and agree with each other

Disagree = Both Rater 1 and Rater 2 comment on the same criteria but disagree with each other and

No Match = Rater 1 and Rater 2 comment on different criteria. For example, Rater 1 may comment on Overall Impression of a particular test but Rater 2 makes no comment in this category. Similarly, Rater 2 may make a comment on Grammar but Rater 1 does not.

9. Results

9.1 To what extent do raters focus on the same criteria when not following a detailed rating schedule?

Although the raters had agreed to focus on three general criteria, the post -hoc analysis of their comments actually identified eight specific criteria. These are listed in Table 1.

Criteria	Rater 1		Rater 2		Total	
1. Overall Impression	17	39%	11	23%	28	31%
2.Content	16	35%	19	40%	35	38%
3.Order of Information	2	4%	2	4%	4	4%
4.Format	3	6%	2	4%	5	5%
5.Sentence Structure	5	12%	1	2%	6	6%
6.Word Choice	0	0%	2	4%	2	2%
7.Grammar	2	4%	7	15%	9	10%
8.Clarity	0	0%	4	8%	4	4%
TOTAL	45	100%	48	100%	93	100%

Table 1. Language Experts' Assessment Criteria (1)

Two categories, *Overall Impression* and *Content* comprise nearly 70% of the total comments made by the two raters. One category, *Overall Impression,* was not even mentioned in the initial discussion yet comprises 31% of the total comments, perhaps indicating a desire by the trainers to look at the tests in a more global manner than the three original criteria allowed. Another category, *Content* (whether the writer had included or omitted relevant information in each section of the report) comprises nearly 40% of the total comments. This large percentage of comments on *Content* is probably related to the prominence given to this topic in the workshop. Participants had seemed very unsure about exactly what information should be included in each section of a report. It would seem that one of the dangers of not using an explicit rating scale is that raters may tend to focus on aspects of writing that have been given prominence in the course and overlook other important aspects of writing.

For statistical purposes, these eight categories were collapsed into four: *Overall Impression, Content, Organisation* and *Linguistic Resources* so that a Chi-square test could be carried out. The Chi-square test produced the following results. (Expected counts occur in italics below the observed counts.)

Criteria	Rater 1	Rater 2	Total
Overall Impression	17	11	28
	13.55	*14.45*	
Content	16	19	35
	19.4	18.06	
Organisation	5	8	13
	6.29	*6.71*	
Linguistic Resources	7	10	17
	8.23	*8.77*	
Total	45	48	93

(Chi^2 =2.671, df=3)

Table 2. Assessment Criteria (2)

The Chi-square result of 2.67 is non-significant. That is, the raters do not differ significantly on the criteria that they mention. This is an interesting result in view of the fact that the raters were not following a detailed, explicit rating schedule. On the contrary, the trainers were making subjective comments on what they each felt to be the most salient points for each test. Thus, it would appear that reliability can be achieved 'by raters agreeing, not necessarily consciously, on criteria for assessment which are only partially explicit in the scoring criteria' (McNamara, 1996:227). McNamara argues that this is 'likely to be a product of the training of the raters as language teachers' (McNamara, 1996:227). I would go further and argue that this could also be a result of experience as trainers, of knowledge of the training context and clientele, and of rapport that comes from trainers working closely together.

9.2 To what extent do the raters agree when using the same criteria?

Although it is heartening to find that the two raters focussed on similar criteria when evaluating the writing tasks, it is also necessary to measure the extent of agreement between raters when using these criteria. The degree of agreement is shown in the following table.

	Comments Match (ie both raters comment on same criteria)		Comments Do Not Match (ie raters do not comment on same criteria)
	Agree (Rater 2)	Disagree (Rater 2)	No Match (Rater 2)
Agree (Rater1)	18 (25%)		
Disagree (Rater 1)		5 (7%)	
No Match (Rater 1)			50 (68%)

Table 3. Extent of Agreement between Language Expert Raters

Somewhat disturbingly, the raters' comments only match up on 32% of the occasions. Fortunately though, where the rater comments match, they agree more than they disagree (25% and 7% respectively). In 68% of the comments there is no match between the raters' comments at all. This high proportion of No Match comments indicates that the raters are not methodically analysing each test for each criterion. They are commenting on only those aspects that seem salient to them. Bachman points out that,

> *the primary causes of inconsistency [in both intra and inter rater reliability] will be either the application of different rating criteria to the different samples or the inconsistent application of the rating criteria to different samples.*
> *(Bachman, 1990:178)*

This result indicates a need for a simple evaluation instrument that would require each rater to systematically work through each criterion for each test. Only in this way could the reliability that seems to be emerging from the results of the Matched comments be tested for significance. If the instrument were to use a 3–5 point scale instead of descriptive comments, the degree of inter-rater reliability could be easily established, thereby enhancing the credibility of the assessment procedures. Such an instrument needs to be both easy to administer and explicit. Although many instruments and scales for assessing writing skills are available (eg IELTS, ASLPR, OET), not all are easy to administer and some are more complex than necessary for the purpose of a 'low-stakes' exercise such as this. The rating instrument developed by the subject expert goes some way towards meeting the practical requirements of such an instrument in that it is easy to administer. However, the four criteria needed to be made more explicit as there was some confusion about precisely what each criterion actually covered.

10. Issue 2: Inter-rater Reliability: Subject Expert and Language Expert

It is usual for the trainer to work closely with management at the needs analysis stage of a training program to identify the training solutions required. It is less common for the trainer and management to work together at the assessment phase of the program.

> *While it is generally accepted that subject specialists should be consulted during the needs analysis phase... their role in the actual assessment process is seldom considered (Elder, 1993:249).*

More involvement in the assessment stage for the subject specialist offers some advantages. It should be remembered that participants in any language skills training course are not attending in order to improve their communication skills with linguists. Outside of the workshop, they need to communicate largely 'with people who have no training in linguistics or language teaching or testing' (Barnwell, 1989:154). Therefore, the subject expert has a valuable contribution to make, given his/her 'insider' knowledge of what is acceptable in the particular profession in which he/she works. If subject specialists and language experts can agree on the criteria for assessment and can both participate in the ranking of participants, then this would further validate the assessment process.

This section of the paper addresses the second issue of inter-rater reliability by exploring the following questions:

1. Do the subject expert and language expert approach the task of assessment in different ways?

2. To what extent do the subject expert and the language expert use the same assessment criteria?

3. To what extent do the subject expert and the language expert's assessments agree?

11. Method

The subject expert (Rater 3) was asked to assess the twelve post-tests. He was puposely given only minimal guidelines on how to assess these tests so that he did not feel constrained to assess them in a pre-determined way. He was encouraged to establish his own assessment framework, as the study was also interested in observing how he approached the whole task of assessment.

Once the subject expert had established his criteria and rated the tests, one of the trainers assessed the post-tests again using the rating schedule that the subject expert had established. Although having one of the trainers reassess the tests was not ideal, three months had elapsed since the initial assessments. Reliability ratings between the subject expert and the language expert were calculated using Pearson's r formula. The results are shown in **Error! Reference source not found.**.

12. Results

12.1 Do subject experts and language experts approach the assessment task differently?

The subject expert approached the assessment task in a quantitative way. He devised a chart with four assessment criteria and a 5-point rating scale, 1= unsatisfactory to 5 = excellent. This is in marked contrast to the two language experts who adopted a more holistic, qualitative approach that relied on brief comments about features they thought salient.

The small number of raters involved in this case study (three) does not allow any generalisations to be made about the differences in approach taken by the language and subject experts. However, if (as is likely in this particular branch of engineering), this quantitative approach were the usual way of dealing with an assessment task it may be prudent for trainers to adopt a similar approach in the interest of improving the credibility of their assessment methodology.

12.2 To what extent do the subject expert's criteria match with the language experts' criteria?

The salient criteria identified by the subject expert were *Structure, Content, English Expression* and *Overall Impression and Style*. It is reassuring to note that these four criteria correspond closely to the criteria used by the two language experts. See Table 4.

Subject Expert	Language Experts
Structure	Organisation (Format, Order of Information, Clarity)
Content (Factual Replication)	Content
English Expression (Grammar, Spelling etc.)	Linguistic Resources (Grammar, Word Choice, Sentence Structure)
Overall Impression & Style	Overall Impression

Table 4. Assessment Criteria Used by Subject and Language Experts

In fact, the two categories of *Overall Impression* and *Content* correspond very closely. However, the subject expert's other two categories of *Structure* and *English Expression* are narrower in scope than the language experts' corresponding categories. For example, *Structure* corresponds (from a personal communication) closely with only one section of *Organisation*, namely *Format*.

Hadden found similar results in her study of teacher and non-teacher perceptions of second language communication, which revealed

> *... a notable similarity in ESL teacher and non-teacher perceptions of second language communication, in this case, of advanced ESL students who are native speakers of Chinese. The dimensions along which the groups evaluated the students, although not identical, were quite similar (Hadden, 1991:17-18).*

12.3 To what extent do the subject and language expert's assessments agree?

The mean of the ratings given by both the subject and the language expert are virtually identical: X=14. The standard deviations are s=2.27 and s=3.24 for the subject expert and the language expert respectively. This data indicates that while the average score of each rater is similar there is considerable variability in the individual scores of each rater . However, the language expert's scores are more widely scattered than the subject expert's scores.

	Mean	Standard Deviation
Tests assessed by Subject Expert (n=12)	14.16	2.27
Tests assessed by Language Expert (n=12)	14	3.24

Table 5. Means and Standard Deviations of Raters

Table 6 illustrates the rankings that the two raters gave the twelve tests. Six out of the twelve rankings differed by more than 4 points. In particular, the rankings of Tests 1 and 8 differed by 8 and 9 points respectively.

Test No.	Rankings		
	Language Expert (Rater 2)	Subject Expert (Rater 3)	Difference
1	9	1	8
2	7	6	1
3	1	6	5
4	6	2	4
5	10	11	1
6	11	9	2
7	12	5	7
8	2	11	9
9	3	2	1
10	5	4	1
11	3	8	5
12	7	9	2

(r= 0.03)

Table 6. Test Ranking of Subject and Language Experts

The correlation between the subject and language experts' rankings is very weak (r =0.03). This is confirmed in a scatter plot of the rankings that indicates that there is virtually no relationship between the two raters (Table 7).

```
   R2
    -
12.0+                       o
    -                                           o
    -
    -                                                     o
    -   o
 8.0+
    -                            o              o
    -        o
    -
    -                  o
 4.0+
    -        o                             o
    -
    -                                                     o
    -                            o
 0.0+
     --------+---------+---------+---------+---------+-----R3
            2.0       4.0       6.0       8.0      10.0
```

Table 7. Scatter plot of Inter-rater Reliability of Subject and Language Experts

This total lack of a relationship is somewhat surprising, even given the limited amount of research conducted on comparisons between subject and language expert rating behaviour. In her review of the research, Hadden notes that

> *Few researchers have compared judgements of nonnative communication by language teachers and nonteachers, and the results of research with this comparative approach have been mixed* (Hadden, 1991:5).

It is interesting to hypothesise why the correlation between the two experts is so low. It could just be due to idiosyncratic marking by both raters, based as it is on only one set of ratings for each rater. However, a close analysis of the editing comments made by both raters for Tests 1 and 8, the tests with the greatest differences between scores, is revealing in that it indicates that although the raters were focusing on similar criteria they were rating the same tests in very different ways.

12.3.1 Post Test 1

A comparison of the editing comments made by each rater for post-test 1 reveals that in fact both raters are highlighting similar errors. Therefore, the problem seems to be in how the raters are scoring these errors. In this particular test, the subject expert, for example, rates more leniently than the language expert, who scored post-test 1 lower on Content, English Expression and Overall Impression. This is in contrast to Brown's study (1995), which found that 'raters with an industry background were found to be harsher than those with a teaching background' (Brown,1995:8). (However, Brown's study also found that although harsher, the differences were non-significant.) In post-test 8, the situation is reversed, with the subject expert rater rating more harshly than the language expert.

		Structure	Content	English	Overall Impression	Total	Rank
Post test 1	Subject Expert (Rater 3)	4	5	5	4	18	1
	Language Expert (Rater 2)	4	4	2	2	12	9
Post Test 8	Subject Expert	2	3	3	3	11	11
	Language Expert	5	5	4	5	19	2

Table 8. Comparison of Ratings for Post-Tests 1 & 8

McNamara identifies differences in 'overall leniency' (McNamara, 1996:125) as one of four ways in which raters may differ from each other in rating. Recent research suggests that 'rater training can reduce but by no means eliminate the extent of rater variability in terms of overall severity' (McNamara, 1996:126).

Where ratings are close, as in Structure (5 and 4), the variation could be accounted for by different interpretations of the rating scale. If unsure about whether the rating should be a 4 or a 5, rater 3 may be tending to score 'up' while rater 2 may be scoring 'down'.

12.3.2 Post-Test 8

A comparison of the editing for this test reveals that the subject expert made many more editing alterations than did the language expert. Consequently, the subject expert scored this test consistently lower on all four categories. What is interesting to note here is that in subsequent discussions, the subject expert stated that headings and numbering were vital to the structure of the report and that his low scores for Structure and Overall Impression were a result of these elements not being included in this post-test. On the other hand, because this was a short memo report, the language expert had not thought that these were relevant and had not penalised the reports for their absence. It seems that the subject and language experts were interpreting these criteria quite differently, each having different ideas about what is important. Also, the subject expert may have been behaving more in line with what McNamara(1990) calls the 'strong' approach to performance testing, which sees performance testing more in terms of task fulfilment than the 'weak' view, that sees performance testing more in terms of the quality of the writing. The differences in interpretation of the criteria highlight the need for very clear guidelines about what each criterion describes.

In hindsight, the two objectives, the first, having the subject expert decide on his own criteria for rating the tests and the second, having him rate the tests, were incompatible and should have been separated. Once the subject expert had established his criteria, they should have been discussed and trialed before the two raters independently rated the tests. If this had been done, I suspect that the correlation between the ratings of the subject and language experts would have been stronger.

It seems that rater training is an important variable when raters from different professional backgrounds are used for assessment. In their study of inter-rater reliability of four groups of raters—language teachers and lay raters, both with and without training, Shohamy, Gordon and Kraemer (1992) found that 'trained raters had higher inter-rater reliability and that teaching background did not make a difference' (cited in Brown 1995: 3).

13. Conclusion

This study has highlighted a number of issues related to inter-rater reliability in the context of measuring achievement through pre- and post-tests. It has been argued that even though the language and subject experts approached the assessment process from different methodological perspectives, they still managed to focus on similar criteria. To ensure that these criteria were addressed in a more systematic way, the language experts would have benefited from using a simple assessment instrument such as the one developed by the subject expert. This would have ensured that each trainer addressed each criterion for each piece of writing.

A detailed analysis of two of the post-tests suggests that while both the subject and language experts were focussing on similar criteria they were actually scoring the criteria quite differently. Rater training could have considerably strengthened the correlations between the two raters. Further research is needed to test Brown's conclusion that

> *... given adequate training and explicit assessment criteria, there is little evidence ... that raters with a teaching background are more suitable than those with an industry background (or vice versa) (Brown, 1995:13).*

Finally, in an ESP context, the subject expert's contribution to the assessment phase of the training cycle may have been underestimated. Further studies are warranted to investigate how their considerable 'insider' knowledge can best be tapped.

14. References

Bachman, Lyle F. 1990 *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press

Barnwell, J. 1989 Naive native speakers and judgements of oral proficiency in Spanish- *Language Testing* 6: 152-63

Brown, Anne 1995 The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12,1: 1-15

Elder, C. 1992 How do subject specialists construe classroom language proficiency? *Melbourne Papers in Language Testing* 1,1: 16-29

Hadden, B. L. 1991 Teacher and non teacher perceptions of second language communication *Language Learning* 41: 1-24

Henning, G. 1987 *A Guide to Language Testing*. Cambridge: Newbury House

Huckin, T. & L.A. Olsen 1991 *Technical Writing and Professional Communication for Nonnative Speakers of English*. New York: McGraw Hill

McNamara, T.M. 1990 Assessing the language proficiency of health professionals. Unpublished PhD thesis, The University of Melbourne

McNamara, T.M. 1996 *Measuring Second Language Proficiency*. London: Longman

Shohamy, E., C.M. Gordon & R. Kraemer 1992 The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal* 76: 27-33.