English Language Teaching
www.ccsenet.org/journal.html
Vol. 2, No. 3
September 2009

The Inter-rater Reliability in Scoring Composition

Ping Wang
School of Foreign Languages, Northwest University of Politics & Law
300, Changan South Road, Xi'an, Shaanxi, China, 710063
E-mail: rainy0329@163.com

Abstract

This paper makes a study of the rater reliability in scoring composition in the test of English as a foreign language (EFL) and focuses on the inter-rater reliability as well as several interactions between raters and the other facets involved (that is examinees, rating criteria and rating methods). Results showed that raters were fairly consistent in their overall ratings. This finding has the great implications for controlling and assuring the quality of the rater-mediated assessment system.

Keywords: Inter-rater reliability, Scoring composition, Rating criteria

For a long term, the experts in language testing are always in dispute about whether the subjective items (for example composition) should be utilized in the crucial tests and the chief objection to the inclusion of the composition as part of any test is generally on grounds of unreliability. Considerable research in the past has shown that unreliable examiners are both in their own inconsistency (intra-rater reliability) and in their failure to agree with colleagues (inter-rater reliability) on the relative merits of rating scale, severity and leniency and so on. In spite of all such demonstrations of unreliability, composition is still widely used in various kinds of language tests merely because it can provide not only a high motivation for writing, but also an excellent backwash effect on teaching. Therefore, if a more reliable means of scoring the composition can be used, sampling a student's language skills by writing will appear a far more valid way than any other objective items.

In this paper the author may concentrate on how to establish high rater reliability, especially the inter-rater reliability in scoring composition. The study is based on a practical research: asking eight examiners to score a composition by using the two different methods (holistic scoring and analytic scoring).

1. The Related Terms

1.1 Reliability

Reliability is the extent to which test scores are consistent: if candidates took the test again after taking it today, would they get the same result. There are several ways of measuring the reliability of "objective" tests (test-retest, parallel form, split-half, KR20, KR21, etc.). The reliability of subjective tests is measured by calculating the reliability of the marking; this is done by several ways (inter-rater reliability, intra-rater reliability, etc.)

1.2 Inter-rater reliability

Inter-rater reliability refers to the degree of similarity between different examiners: can two or more examiners, without influencing one another, give the same marks to the same set of scripts (contrast with intra-rater reliability).

1.3 Holistic scoring

Holistic scoring is a type of rating where examiners are asked not to pay too much attention to any one aspect of a candidate's performance, but rather to judge general writing ability rather than to make separate judgement about a candidate's organization, grammar, spelling, etc.

1.4 Analytic scoring

Analytic scoring is a type of rating scale where a candidate's performance (for example in writing) is analyzed in terms of various components (for example organization, grammar, spelling, etc.) and descriptions are given at different levels for each component.

2. The Methods Used to Obtain High Inter-rater Reliability

2.1 The Importance of High Inter-rater Reliability

In common sense, it would not be realistic to expect all examiners to match the "standard" all the time because if the marking of a test is not valid and reliable, then all of the other work undertaken earlier to construct a "quality" instrument will have been a waster of time. No matter how well specifications of a test reflect the goals of the institution or how much care has been taken in the designing and protesting of items, all the effort will have been in vain if the test

users cannot have faith in the marks that examiners give to the candidates.

In one word, the poor inter rater consistency will directly reduce the reliability and the validity of the test to a very large degree.

2.2 How to Establish High Inter-rater Reliability

2.2.1 Setting the Standard

In a test with a large number of examinees, it is impossible for all the examiners to have an equal say in determining scoring policy. This description assumes that there is a "Chief Examiner (CE)", who, either alone or with a small group of colleagues, setting the standards for marking and passes these onto the examinees who may mark centrally or individually in their homes.

2.2.2 Training the Scorers

The scoring of compositions shouldn't be assigned to anyone who has not learned to score accurately compositions from past administrations. After each administration, patterns of scoring should be analyzed. The individuals whose scorings deviate markedly and inconsistently from the norm should not be used again.

2.2.3 Identifying Candidates by Number, Not Name

Scorers inevitably have expectations of candidates that they know, this will affect the way that they score, especially in subjective marking. Studies have shown that even where the candidates are unknown to the scorers, the name on scripts will make a significant difference to the scores given. For example, a scorer may be influenced by the gender or nationality of a name into making predictions which can affect the score given. The identification of the candidates only by number will reduce such effects.

2.2.4 Setting the Specific Standards before the "Real Scoring"

So after the test has been administered, the CE should read quickly through as many scripts as possible to extract scripts which represent "adequate" and "inadequate" performances, as well as scripts which present problems which examiners are often faced with but which are rarely described in rating scales: bad handwriting, excessively short or long responses, responses which indicate that the candidates misunderstood the task etc.

The next step is for CE to form a standardizing committee to try out the rating scale on these scripts and to set and record the standards. All of the marking members should be given copies of the scripts selected by the CE, in random order, and each member should mark all of these scripts before the committee meets to set standards.

2.2.5 Sampling by the Chief Examiner or Team Leader

Each examiner is expected to make a certain number of scripts on the first day of marking. The team leader collects a percentage of marked scripts from the examiners (often 10-20%), and reads through them again in order to give an independent mark (that is called "blind marking") to find whether the examiners marking properly. The process of sampling should be continued throughout the marking period in order to narrow the differences in examinees.

2.2.6 Using "reliability scripts"

The second method of monitoring marking is to ask each examiner independently to mark the same packet of "the reliability scripts" which have been marked by the standardizing committee earlier. The reliability exercise should take place after the examiners have begun marking "for real", but early enough in the marking period for changes to be made to scripts which may already have been marked incorrectly by unreliable examiners. The afternoon of the first day of marking or the second morning would be suitable times.

2.2.7 Routine double marking

The third way of monitoring examiners and ensuring that their marks are reliable is to require routine double marking for every part of the exam that requires a subjective judgement. This means that every composition should be marked by two different examiners, each working independently. The mark that the candidate receives for a piece of writing is the mean of the marks given by the two examiners.

3. The Two Ways of Scoring Composition

So far in part II, we have been concerned to improve the inter-rater reliability. Now we'd like to turn to the methods of scoring.

Composition may be scored according to two different criteria: the holistic scoring and the analytic scoring.

3.1 Holistic Scoring

Holistic scoring is a type of rating where examiners are asked not to pay too much attention to any one aspect of a candidate's performance, but rather to judge general writing ability rather than to make separate judgement about a candidate's organization, grammar, spelling, etc. This kind of scoring has the advantage of being very rapid.

Experienced scorers can judge a one-page of writing in just several minutes or even less. As it is possible for each composition to appear just to a certain rater but not others, the examiner's mark may be a highly subjective one. However, if assessment is based on several judgements, the net result is far more reliable than a mark based on a single judgement.

Because the inherent unreliability in holistic marking of compositions, it is essential to combine a banding system, or, at least, a brief description of the various grades of achievement expected to be attained by the examinees. An example of a holistic scale is given in the coming figure.

Insert Figure 1 Here

3.2 Analytic scoring

Since most teachers have little opportunity to enlist the services of two or three colleagues in marking compositions, the analytic method--analytic scoring--is recommended for such purposes.

Analytic scoring is a type of rating scale where a candidate's performance (for example in writing) is analyzed in terms of various components (for example organization, grammar, spelling, etc.) and descriptions are given at different levels for each component (see Figure 2).

Insert Figure 2 Here

These rating criteria (Figure 1 and Figure 2) are only two of many that are available in EFL testing. The number of points on the scale and the number of components that are to be analyzed will vary, given the distinct demands that different writing tasks can place on candidates. The challenge to examiners is to understand the principles behind the particular rating scales they must work with, and to be able to interpret their descriptors consistently.

4. Give a Composition for Eight Examiners to Score

In order to make the study, the author of this paper chose one composition from the examinees' works and eight examiners to mark the composition individually. The raters who marked the examinee's writing were all experienced teachers and specialists in the field of English as a foreign language. Each rater was licensed upon fulfillment of strict selection criteria. As mentioned previously, raters were systematically trained and monitored as to compliance with scoring guidelines. Ratings of examinee's essay were carried out according to the two main marking methods mentioned previously: holistic marking method and the analytic marking method. The analytic marking method includes a detailed catalogue of performance aspects: content, organization, cohesion, vocabulary, grammar, punctuation and spelling etc. (for the detailed information, please consult Figure 1 and Figure 2 in Part3).

The coming table shows us the scores given by the eight examiners, including holistic marking scores and the sum scores of each candidate according to the analytic scales.

Insert Table 1 Here

5. Data Analysis

And then the author analyzes the scores with the help of the statistical software SPSS and gets the statistic data (presenting below).

M1 is the marks given by holistic scoring

Mean of M1=12.3750, range =5, SD=1.50594, SD error mean=.53243

M2 is the marks given by analytic scoring

Mean of M2=14.5000, range=5, SD=2.07020, SD error mean=.73193

The correlation between M1 and M2 is .802; significance level .017 (consulting Table2 and Table3).

F of M1 is 1.188 and F of M2 1.705, both of them are less than df, 5 and 4 individually (consulting Table4 and Table 5) so we can get the conclusion that the differences among the eight examiners are not obvious.

Insert Table 2, Table 3, Table 4, Table 5 Here

6. Conclusions

In this paper the author first showed the importance of high inter-rater reliability in EFL testing and told us how to gain the high inter-rater reliability (there are seven ways mentioned in this paper). Then the author tried to determine whether the raters are consistent in scoring the subjective items (taking composition as an example) by using the different scoring methods (holistic scoring and analytic scoring). At the end of the paper the author, by analyzing the data, got the conclusion that raters were fairly consistent in their overall ratings (the correlation is .802, significance level is .017) and the marks given by analytic scoring are usually a little higher than that of holistic scoring (mean of M1 is 2.1250 less than mean of M2). This finding has the great implications for controlling and assuring the quality of the rater-mediated assessment system.

References

J. Charles Alderson, Caroline Claphan and Diame Wall. (2000). *Language Test Construction and Evaluation*.Beijing*:* Foreign Language Teaching and Research Press.

Li, Shaoshan. *Basic Statistics in Language Studies.*Shaanxi: Xi'an Jiaotong University Press.

Li, Xiaoju. (1997). *The Science and Art of Language Studies* Hunan Educational Press.

Lyle F. Bachman *Fundamental Considerations in Language Testing* Shanghai Foreign Language Educational Press.

Shephard, L.A. (1984). Setting Performance Standards in Berk.

Shohamy, E. and T. Reves. (1985). Authentic Language Tests: Where from and Where to? Language Testing.

Shohamy, E. (1983). Inter-rater and intra-rater Reliability of the Oral Interview and Concurrent Validity with Close Procedure in Hebrew.

Table 1. The Marks Given by the Teachers

	Holistic Scoring Marks(M1)	Analytic scoring marks							Total Marks (M2)
		Content	Organization	Cohesion	vocabulary	Grammar	Punctuation	Spelling	
Teacher1	15	3	2	2	2	3	2	3	17
Teacher2	13	2	2	2	2	3	3	3	17
Teacher 3	10	2	2	2	1	1	2	2	12
Teacher 4	12	2	2	2	2	2	2	2	15
Teacher 5	13	3	2	2	2	3	2	2	16
Teacher 6	11	2	2	2	2	2	2	2	13
Teacher 7	13	2	2	2	2	2	2	3	14
Teacher 8	12	2	2	1	2	2	2	2	12

Table 2. Paired Samples Statistics

	N	Minimum	maximum	Range	Mean	Std. Deviation	Std. Error Mean
.M1	8	10	15	5	12.3750	1.50594	.53243
M2	8	12	17	5	14.5000	2.07020	.73193

Table 3. Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	M1 and M2	8	.802	.017

Table 4**.** ANOVA (M1)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11.875	5	2.375	1.188	.516
Within Groups	4.000	2	2.000		
Total	15.875	7			

Table 5. ANOVA (M2)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20.833	4	5.208	1.705	.345
Within Groups	9.167	3	3.056		
Total	30.000	7			

18-20	Excellent	Natural English with minimal errors and complete realization of the task set
16-17	Very good	More than a collection of simple sentences, with good vocabulary and structures. Some non-basic errors.
12-15	Good	Simple but accurate realization of the task set with sufficient naturalness of English and not many errors.
8-11	Pass	Reasonably correct but awkward and non-communicating or fair and natural treatment of subject, with some serious errors.
5-7	Weak	Original vocabulary and grammar both inadequate to the subject.
0--4	Very poor	Incoherent. Errors show lack of basic knowledge of English.

Figure 1. A Sample Holistic Scale

From: UCLES International Examinations in English as a Foreign Language General Handbook, 1987

Relevance and Adequacy of content

0. The answer bears almost no relation to the task set. Totally inadequate answer.

1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.

2. For the most part answers the task set, though there may be some gaps or redundant information.

3. Relevant and adequate answer to the task set.

Compositional Organization

0. No apparent organization of content.

1. Very little organization of content. Underlying structures not sufficiently apparent.

2. Some organization skills in evidence but not adequately controlled.

3. Overall shape and internal pattern clear. Organization skills adequately controlled.

Cohesion

0. Cohesion almost totally absent. Writing is so fragmentary that comprehension of the intended communication is virtually impossible.

1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.

2. For the most part satisfactory cohesion though occasional deficiencies may mean that certain parts of communication are not always effective,

3. Some use of cohesion resulting in effective communication.

Adequacy of Vocabulary for Purpose

0. Vocabulary inadequate even for the most basic parts of the intended communication.

1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetitions.

2. Some inappropriacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.

3. Almost no inappropriacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

Grammar

0. Almost all grammatical patterns inaccurate.

1. Frequent grammatical inaccuracies.

2. Some grammatical inaccuracies.

3. Almost no grammatical inaccuracies.

Mechanical Accuracy I (Punctuation)

0. Ignorance of conventions of punctuation.

1. Low standard of accuracy of punctuation.

2. Some inaccuracies of punctuation.

3. Almost no inaccuracies of punctuation.

Mechanical Accuracy II (Spelling)

0. Almost all spelling inaccurate.

1. Low standard of accuracy in spelling.

2. Some inaccuracies in spelling.

3. Almost no inaccuracies in spelling.

Figure 2. A Sample Analytic Scale

From: *Test of English for Educational Purposes*, Associated Examining Board, UK, 1984.