Revista de Lenguas Modernas, N° 20, 2014 / 295-307 / ISSN: 1659-1933

Intra-rater Reliability and the Role of Experience: A Comparative Case

Isabel Cristina Bolaños Villalobos
Gabriela Cerdas Ramírez
Jimmy Ramírez Acosta
Escuela de Literatura y Ciencias del Lenguaje
Universidad Nacional, Costa Rica

Abstract
This article aims to present the results of an investigation which seeks to determine whether experience plays an important role in increasing intra-rater reliability. The paper also provides an analysis that is carried out to determine if factors such as gender or rubric development affect raters' reliability. Finally, a list of practical implications drawn from this investigation is provided.

Key words: rubric, assessment, rater reliability, Pearson product-moment correlation coefficient

Resumen
Este artículo presenta los resultados de una investigación en la cual se pretende determinar si la experiencia juega un papel importante en el aumento del nivel de confiabilidad de evaluación de un grupo de profesores universitarios. Además, se realiza un análisis para determinar si aspectos como el género o el desarrollo de una escala de evaluación tienen incidencia en la confiabilidad de los evaluadores. Finalmente, se realiza un apartado con algunas implicaciones prácticas desprendidas de esta investigación.

Palabras claves: rúbrica, evaluación, confiabilidad del evaluador, coeficiente de correlación producto momento de Pearson

Recepción: 3-3-14 Aceptación: 23-4-14

Since students are expected to succeed in their process of learning a language, it is not surprising that professors are not only required to do their best when teaching their classes, but also when assessing their students. This reasoning may put too much pressure on professors since they have to grade students and at the same time comply with high reliability levels. Many EFL professionals will agree that professors in general should concentrate on their ability to come up with reliable grades due to the pedagogical effect this has on the learning process. When professors are reliable raters, the learning process is enhanced because the learners and the professor can make decisions to improve the acquisition process of the target language. The professor will be able to identify true weaknesses and strengths of the students' performance.

There are studies in which investigators have come up with different and useful findings about rater reliability but with a very homogeneous sample; for example, Hayes and Hatch proposed the use of correlation measures to determine reliability,[1] and Cohen who proposed the use of kappa coefficient to determine rater reliability.[2] There are not, however, studies on intra-rater reliability in which experience is considered as an important influencing variable. This research is intended to determine whether experience plays an important role on the teachers' capacity to score students appropriately and consistently over time.

What is reliability?

Moskal defines reliability as the consistency of scores that are assigned by two independent raters or that are assigned by the same rater at different points in time.[3] In addition, it is important to indicate that some authors have also proposed the concept of *True Score Theory,* Brogan; for example, indicates that *True Score* is the exact measure of the test taker's true ability in the area being tested. With a perfect test, the observed score would be equal to the true score.[4] Based on what Brogan proposes, it is clear then that when there are changes in the grades of the students, this must be due only to changes in performance or in the ability of the person evaluated. Changes in students' evaluations should never be the result of the rater's ability to generate reliable grades.

Rubrics

One of the conditions that might help professors become reliable raters is the use of rubrics. Johnson, for example, indicates that professors are likely to improve the way they grade students' output when they use an appropriate rubric.[5] Hafner gives a very simple but accurate definition of rubric which is relevant for the purpose of this investigation. According to Hafner, "In the educational literature and among the teaching and learning practitioners, the word 'rubric' is understood generally to connote a simple assessment tool that

describes levels of performance on a particular task and is used to assess outcomes in a variety of performance-based contexts." [6]

Holistic versus analytic rubrics

It is clear to many EFL professionals that they cannot even dare to grade students' output without using a proper rubric. This might be the reason why there are holistic as well as analytic rubrics which are commonly used in the EFL environment. Johnson, for example, describes holistic rubrics as those that are used when the rater makes an overall judgment about the quality of performance. On the other hand, analytic rubrics are used when the rater assigns a score to each of the dimensions being assessed in the task [7]. In Costa Rica, both types of rubrics are known and common; however, professors tend to use more analytic rubrics over holistic ones. This is surprising since both types of rubrics allow the teachers and the students to make significant decisions because they can have a clear idea of what their strengths and learning needs are.

Since analytic rubrics have such an important role in the learning process, the need of reliability becomes imperative. It is important to acknowledge that reliability is divided into two different types: inter-rater reliability, and intra-rater reliability. Inter-rater reliability has to do with differences in grades that are obtained from different raters. This will happen, for example, when a student is being interviewed, and three different professors, using the same rubric, come up with three different grades. On the other hand, intra-rater reliability has to do with differences in grades obtained from one single rater; for example, a professor who is grading one student and comes up with one grade at a specific moment, but this professional comes up with a different grade for the same student with the same performance on another day. The previous case is of great interest due to the pedagogical consequences it might have on the students' development. Some professionals might accept that there is a chance that raters assign different grades and this, according to Johnson, might be due to differences in experience, lack of agreed-upon scoring routines, teachers' attitudes regarding students' ethnicity, as well as the content.[8] The problem with this issue is that the subjects involved in the educational process would not feel comfortable to trust the results obtained in a determined task because they are likely to change from one day to another.

The Pearson product-moment correlation coefficient

Being able to accurately determine someone's intra-rater reliability is an important process any EFL professional must strive to achieve. According to Brown, "Intrarater reliability is typically estimated by getting two sets of scores produced by the same rater for the same group of students, and calculating a correlation coefficient between those two sets of scores."[9] Brown proposes the use of

the Pearson product-moment correlation coefficient to estimate the consistency of judgments made, over time, by the same rater.

The Pearson product-moment correlation coefficient, also known as the regression coefficient, assesses how well the relationship between two variables can be described. This is the formula that is commonly used to determine the correlation

$$r_{xy} = \frac{\Sigma (X-M_x)(Y-M_y)}{NS_xS_y}$$

Where

r_{xy} = Pearson product-moment correlation coefficient
X = The grades given by the professor to each one of the students in the first phase of the investigation
Mx = Mean of the grades obtained in the first phase of the investigation
Sx = Standard deviation of the grades obtained in the first phase of the investigation
Y = The grades given by the professor to each one of the students in the second phase of the investigation
My = Mean of the grades obtained in the second phase of the investigation
Sy = Standard deviation of the grades obtained in the second phase of the investigation
N = Number of students who were evaluated

Why should professors become reliable raters?

Since professors as well as students are concerned about ways to improve the teaching and learning process, being able to come up with reliable grades is imperative. These are some of the reasons why professors should strive to become reliable raters:

1. Professors who have a coefficient of at least +0.80 will know that the grades they assign are consistent. This means that these grades are not likely to vary through time if the students' performance does not vary.
2. It is clear that based on the students' grades, decisions can be made in order to overcome possible learning limitations. This means that if the scores the professors assign are reliable, the students and the professor can have a clear perspective of those aspects that really need to be improved.

3. Another aspect that is highly enhanced when professors are reliable raters is transparency. Students are not likely to complain about the grades they get because they believe in the professors capacity to assign grades that truly match their real performance.
4. If the grades obtained are reliable, professors are likely to measure the effectiveness of their teaching practices. They can clearly see if what students are intended to know has been properly learned.

Methodology

In order to carry out this investigation, 20 university professors were chosen, and they were asked to provide some information about their teaching experience and their rubric development.

Since the main objective of this investigation is to determine whether experience plays a role on intra-rater reliability, professors were classified as novice and experienced. For the purpose of this investigation, professors who had been teaching for less than five years were classified as novice.

Since the Pearson product-moment correlation coefficient is going to be used to determine intra-rater reliability, it is necessary to carry out two different procedures. In the first stage, the professors were asked about their teaching experience and their rubric development. In this stage, they were also given a file with some recordings of students speaking English. The professors were asked to grade the students based on their performance in an oral presentation.

In the second stage and due to the fact that intra-rater reliability has to do with the "consistency of scores that are assigned by two independent raters or that are assigned by the same rater at different points in time,"[10] the subjects were asked to grade the students' performance again one month after they had handed in their first set of grades.

For each of the professors who took part in this investigation, a table with the grades assigned to each student, the mean, the standard deviation, and the range was completed in order to obtain the correlation coefficient.

Table 1
Information needed to obtain a correlation coefficient

1	2 X	3 M_X	4 $(X-M_X)$	5 Y	6 M_y	7 $(Y-M_y)$	8 $(X-M_X)$ $(Y-M_Y)$
1							
2							
3							
4							

9	N= M= S= Rang	

1= Number assigned to each student in order to be rated.

2= The grades given to each student in the first stage of the investigation will be presented in order from the highest to the lowest.

3= The mean of all the grades assigned by the professor.

4= It is necessary to subtract the mean from each of the scores obtained in the first stage of the investigation.

5= The grades obtained by the students in the second stage of the investigation. The grades will not be ordered from the highest to the lowest because each pair of scores is independent.

6= The mean of all the grades assigned by the professor in the second stage of the investigation.

7= It is necessary to subtract the mean from each of the scores obtained in the second stage of the investigation.

8= Results from column 4 and 7 multiplied times each other for each of the students.

9= N= number of participants.
M= mean obtained in each stage of the investigation.
S= standard deviation obtained in each stage of the investigation.
Range= number of points between the highest and the lowest score plus 1.

All this information will be calculated through the use of Pearson product-moment correlation coefficient which is a means of determining if the professors are reliable raters. The results will range from -1 up to +1. According to Brown, a perfect relationship; the one in which the two sets of scores match perfectly, will take the maximum value of +1.0; however, "real scores seldom line up perfectly."[11] If a correlation coefficient of -1 is obtained, there is not a relationship between the two sets of scores. This will be the worst-case scenario because this means that the professor is not a reliable rater at all. "Coefficients either positive or negative up to about +0.40 or -0.40 indicate fairly weak relation. Relatively strong correlations would be those ranging from +0.80 to +1.0, or -0.80 to -1.0."[12] Thereby, for this investigation, a correlation of + 0.80[13] is enough to conclude that the subjects of the study are reliable raters.

Data analysis

The information will be analyzed in two different ways. First of all, the information that has been obtained from each of the professors who took part in this research will be evaluated and carefully analyzed. In the second part, the results obtained from each of the professors will be compared to the results

obtained from the rest of the professors. By doing this, it will be possible to determine whether experience plays an important role on rater reliability.

The information in the following table refers to one of the subjects who took part in this investigation; the data in columns 2, 3 and 4 refers to the first stage of the investigation whereas the information in columns 5, 6, and 7 refers to the second stage.

Table 2
Information obtained from one of the subjects of the investigation

S	X	M_X	$(X-M_X)$	Y	M_y	$(Y-M_y)$	$(X-M_X)(Y-M_Y)$
3	96	77	19	96	82	14	266
7	92	77	15	92	82	10	150
4	80	77	3	88	82	6	18
5	80	77	3	80	82	-2	-6
8	80	77	3	88	82	6	18
2	72	77	-5	76	82	-6	30
10	70	77	-7	72	82	-10	70
1	68	77	-9	80	82	-2	18
6	68	77	-9	72	82	-10	90
9	64	77	-13	76	82	-6	78
					Σ (X-MX)(Y-MY)		732
	N=	10			N=	10	
	M=	77			M=	82	
	S=	10.08			S=	8.04	
	Range=	33			Range=	21	

$$r_{xy} = \frac{\Sigma (X-M_x)(Y-M_y)}{NS_xS_y} = \frac{732}{10\ (10.08)\ (8.04)} = \frac{732}{810} = 0.90$$

The following table has part of the information that was obtained from each of the participants of this investigation. They were asked about their teaching experience, rubric development and rubric's use in order to determine if these aspects play a role on intra-rater reliability.

Table 3
Information obtained from all the subjects

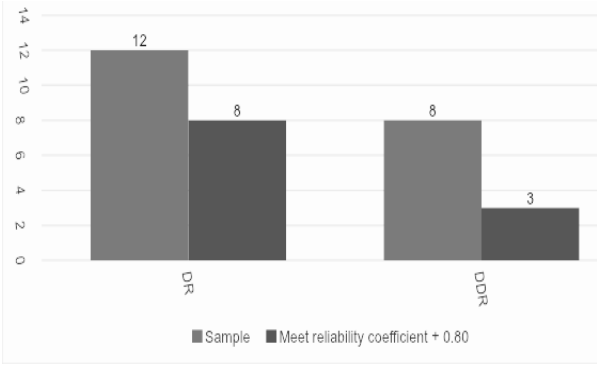
Subject	Gender	TE	RD	RU	ME	CC
15	Male	3	No	1	80.5	0.60

3	Female	12	Yes	7	65.5	0.64
18	Male	1	No	1	88	0.65
2	Male	14	Yes	14	71	0.67
11	Female	4	No	1	75	0.68
14	Male	4	No	1	77	0.69
10	Female	20	No	4	68	0.70
12	Female	3	Yes	1	85	0.75
19	Female	4	Yes	2	81.5	0.76
4	Male	8	No	1	77	0.80
13	Male	5	Yes	3	76	0.80
5	Female	10	No	2	71	0.83
8	Female	6	Yes	1	89	0.85
16	Female	4	Yes	3	88	0.87
6	Female	18	No	7	60.5	0.88
20	Male	5	Yes	3	83	0.88
1	Female	10	Yes	3	73	0.90
17	Female	5	Yes	4	90	0.90
9	Male	25	Yes	11	68	0.91
7	Male	10	Yes	6	72	0.92

TE = Teaching experience.
RD = Rubric development. In this case, the subjects were asked if they had designed the rubric they use.
RU = Rubric's use. In this case, the subjects were asked how long they had used the rubric.
ME = Average mean. It was obtained from the two phases of the investigation.
CC = Correlation coefficient.

When talking about rater reliability, it seems difficult to determine a factor that might help professionals become reliable raters; however, rubric development seems to be an aspect that might help teachers become reliable raters because they would actually know what aspects of the students' performance they would evaluate. Besides, when professors design their own rubric scales, they are likely to know how to assign a score based on the students' performance. The following graph represents the relation between rubric development and intra-rater reliability.

Graph 1
Relation between rubric development and intra-rater reliability



67% of the professors who developed their own rating scale (DR) obtained to be considered reliable raters. This argument supports what most professionals have proven over time, when you develop your own rubrics, you are a more reliable rater. From the information obtained in this research, 38% of the professors who did not develop their own scales (DDR) obtained the appropriate correlation coefficient to be considered reliable raters. Nevertheless, it is important to analyze some of the reasons why people can become reliable raters even when they have not developed their own scales. Some hypothesis to prove this might be:

1. "A scoring rubric with well-defined score categories should assist in maintaining consistent scoring regardless of who the rater is or when the rating is completed."[14]
2. Raters might have used a two level scale which is easier to use because the rater has to determine whether a certain aspect is met or not; however, a four or five-level scale is more difficult to use since raters might have a broader scope of aspects to consider.
3. Some professors use rating scales that are implemented in the institutions where they work. Sometimes these professionals are trained to use the rubric scale appropriately and this might enhance reliability.
4. Some of these professionals might have a deep understanding of the scale they are using, and this would allow them to pay attention to the students' performance and come up with a grade that perfectly matches what the students have done.
5. Unfortunately, there are professionals who might score students based on the number of mistakes the students have made rather than on the type of mistake. They can easily come up with a grade just by looking at the number of mistakes.
6. Sometimes when teachers do not know who they are grading, reliability can be improved. For example, Barbara Moskal indicates that "A correct response from a failing student may be more critically analyzed than an

identical response from a student who is known to perform well."[15] For the purpose of this study, this reasoning applies since the professors did not know the students they were grading.

Even when all the professionals who were part of this investigation believed that the scales they used were reliable and would allow them to come up with reliable scores, 45% of them did not reach the minimum coefficient to be considered reliable raters. In this case, it is difficult to determine if the problems with reliability are due to the rating scale or to the rater. The following are some aspects that would not allow teachers to come up with reliable scores:

1. The scales that the teachers are using do not match the current proficiency level of the students who are being assessed. In order to enhance reliability, the grading scales must match not only the type of activity that is being developed but also the objectives as well as the student's level. If all those aspects are not met, scores are not likely to be valid even if the professor is experienced.
2. The teachers did not have the opportunity to receive training for the rubric's use. Even for the simplest grading scale, teachers need training to learn how to use it. Teachers also need to learn about the principles underlying the development of any rubric, so that they can actually come up with grades that truly match the students' performance.
3. The teachers did not use a rubric scale, and according to Jonsson "Results from studies investigating intra-rater reliability indicate that rubrics seem to aid raters in achieving high internal consistency when scoring performance tasks."[16] It is important to indicate, however, that this reasoning does not apply to this study since professors used a rubric when scoring students' performance.
4. Teachers are tired or in a bad mood at the moment of grading students. If teachers do not feel comfortable when grading students, grades are not likely to reflect student's real performance. It would be difficult to determine whether this variable played an important role in this investigation since this study did not include a process to evaluate or determine the raters' mood.

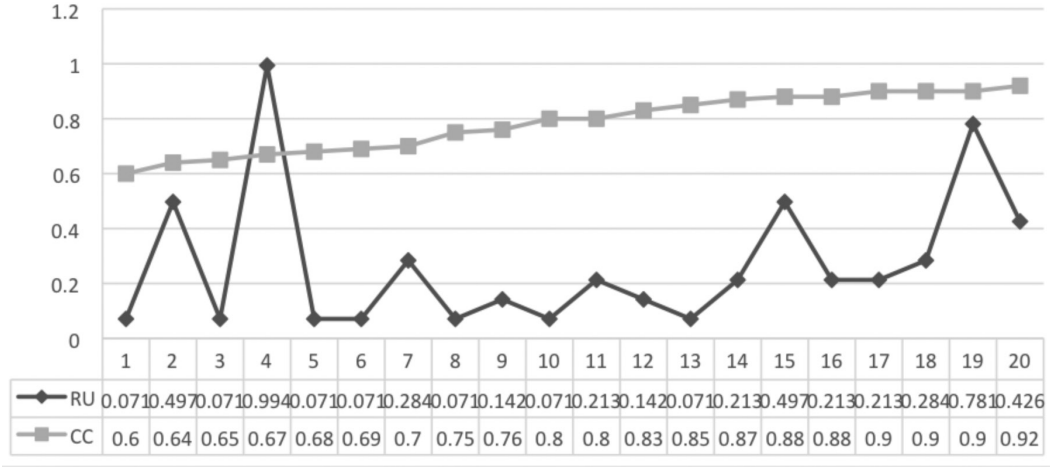
It is believed that experience plays a major role in developing reliability and this can be easily supported by the results obtained in this investigation where there is a clear link between experience and intra-rater reliability. This is not a surprising result since Davidson, for example, indicates that experience is one of the most obvious reasons for differences in grading.[17]

The following graph represents the relationship between the teachers' experience and the correlation coefficient. 63% of the people who obtained an appropriate correlation coefficient in this study were classified as experienced professors. One reason for this might be that experienced professors tend to evaluate oral presentations in an integrated manner.

Graph 2
Relation between experience and correlation coefficient

Most EFL professionals believe that if you have experience in using a rubric, you are likely to come up with more reliable scores; however, based on this investigation, it can be concluded that the experience in the use of a given rubric does not play a major role in the reliability of the scores.

Graph 3
Relation between rubric use and correlation coefficient



When the results are analyzed on the basis of gender, there are no surprises to what Bell and Greatorex had previously concluded: "Sex and gender bias in marking is something which should be monitored … but it is unlikely to be found

to an extent that affects grades."[18] The results obtained from this investigation show that there are no differences between the scores assigned by women and men. The professor´s gender does not play a role in the reliability of the scores these professionals came up with.

Bell and Greatorex concluded from their study that "Experienced examiners are sometimes found to be more lenient than inexperienced examiners;"[19] however, based on the information obtained from this investigation the most experienced professors were the ones who assigned the lowest scores.

This might be due to the fact that experienced professors tend to pay attention to a wider scope of aspects when they grade an oral presentation. Experienced professors do not only pay attention to aspects such as grammar, pronunciation and vocabulary, but they actually analyze aspects such as fluency, communication, task, platform techniques, body language, etc.

Conclusions and practical implications

When professors are able to determine their intra-rater reliability coefficient, they are more likely to improve the learning process since they would be able to analyze if the students have actually learned what they are intended to learn. In addition, the use of a rubric is highly recommended to improve intra-rater reliability, given that rubrics allow raters to focus on specific aspects of the students' performance; however, using rubrics to improve intra-rater reliability does not necessarily mean that the professors have to design them. Intra-rater reliability is improved when the scales that are used truly match the students' level as well as the characteristics of the activity. It is also important to point out that when a professor design a rubric does not necessarily mean the professor is going to become a reliable rater. On the other hand, when professors do not develop their own rating scale, they should be trained on its use because this is a fundamental factor in attaining reliability.

When analyzing gender as a possible influencing factor in reliability, it can be concluded that it does not play a major role in the reliability of scores. Grades might be affected by an endless number of emotional, linguistical, psychological, and environmental aspects, but they are not likely to be affected by gender.

Finally, experience is an aspect that should be carefully analyzed since experienced professors are more likely to pay attention to a broader set of aspects when grading students' performance, and this might be the reason why they tend to be stricter than less experienced professors.

Bibliographical Notes

1 John R. Hayes & Jill A. Hatch, Issues in Measuring Reliability, *Written Communication*, 16, 3 (1999): 354.

2 Jacob Cojen, A Coefficient of Agreement from Nominal Scales, *Education and Psychological Measurements*, 20 (1960): 37.
3 Barbara Moskal & Jon A. Leydens, Scoring rubric development: validity and reliability (2000), <http://PAREonline.net/getvn.asp?v=7&n=10> , accessed April 9, 2012.
4 Ray Brogan, Reliability (2009), <http://www.education.com/reference/article/reliability/>, accessed April 9, 2012.
5 Anders Jonsson, The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences, *Educational Research Review* 2, (2007):130.
6 John C. Hafner & Patti M Hafner, Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating, *International journal of science education*, 25, (2003): 1509–1528.
7 Jonsson, 131.
8 Jonsson, 131.
9 James Dean Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment* (New York, McGraw-Hill ESL-ELT 2005) 187.
10 Moskal.
11 Dean, 142.
12 Dean, 141.
13 Sheryl Ward, Inter-rater Reliability in an ESP Context, *Melbourne Papers in Language Testing,* 6, 1 (1997): 73.
14 Barbara Moskal & Jon A. Leydens, Scoring rubric development: validity and reliability (2000), <http://PAREonline.net/getvn.asp?v=7&n=10> , accessed April 9, 2012.
15 Moskal.
16 Jonsson, 135.
17 Marcia Davidson, Kenneth Howell & Patty Hoekema, Effects of ethnicity and violent content on rubric scores in writing samples, *Journal of Educational Research 93*, (2000): 369.
18 John Bell & Jackie Greatorex, *Does the gender of examiners influence their marking?*, <http://www.cambridgeassessment.org.uk/ca/digitalAssets/113784_Does_the_Gender_of_Examiners_Influence_Their_Marking.pdf >, accessed August 18,2012.
19 Bell.