ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 4, No. 4, pp. 649-654, July 2013

doi:10.4304/jltr.4.4.649-654

L2 Writing Assessment in the Greek School of Foreign Languages

Despoina Panou
Department of Education, University of Leicester, UK

***Abstract*—In the past two decades, there has been an increased interest in the assessment of L2 writing since the results of such evaluations are used for a variety of administrative, instructional and research purposes. One of the primary issues pertaining the assessment of writing quality is the type of scoring procedure which will be used; admittedly a subject of a great deal of research and discussion in the language testing literature. The aim of the present paper is firstly, to briefly describe the type of scoring used in the Greek School of Foreign Languages for assessing L2 writing performance and secondly, to calculate the inter-rater reliability of five written samples. The results obtained indicate quite high correlations, thus demonstating the evaluators' uniformity in the application of assessment criteria.**

***Index Terms*—L2 writing assessment, scoring procedure, inter-rater reliability**

I. Introduction

In the last twenty years there have been developments in testing reflected in the publications now on the market (Bachman, 1990; Fulcher, 2010; Fulcher and Davidson, 2007; O'Sullivan, 2011) and by the arrival of journals such as *Language Testing* and *Assessing Writing*. Irrespectively of the theoretical approach adopted by each scholar, there is a common consensus that testing is not situated solely within the classroom but should be viewed as a social phenomenon, thus being placed in a larger educational context. Guidance on how to conduct assessment and to build good knowledge tests implies knowledge of the larger socio-cultural context of both language learners and institutions. Of the many issues involved in language testing, and in particular, second-language assessment, reliability has been of particular concern to both language testers and teachers.

II. Literature Review

A fundamental concept in language testing is reliability in the sense that in order for a test to achieve its intended purpose, the test results must be reliable (Fulcher 1997, 1999). According to Jones (1979) one factor that significantly influences reliability in a performance test is the consistency of ratings. In fact, four types of reliability have been proposed: (1) inter-examiner reliability, (2) intra-examiner reliability, (3) inter-rater reliability, and (4) intra-rater reliability.This section focuses on the issue of inter-rater reliability. According to McNamara (1996) the issue of reliability tends to be more complicated because of the subjectivity of human raters in assessing a particular test. In other words, no matter how unequivocal the set of criteria by which the test is to be judged, it could be the case that there are differences or even disagreements in the ratings allotted to the same set of tests by a group of human raters. In fact, Hamp-Lyons (1990) argues that the reliability of rating largely depends on the rater's attitudes and conceptions. In this respect, inter-rater reliability is meant to refer to the consistency between two or more raters who assess the same language test (Lombard et al. 2005; Gamaroff, 2000). Thus, inter-rater reliability enables the raters to standardize their scoring and to maintain a check on the level of consistency of their marking. It should be noted that Tinsley and Weiss (2000, p.98) prefer the term 'inter-rater agreement' but here, the term inter-rater reliability will be used referring to its widely accepted sense, that is, the correlation between two rating sets. In this sense, rater-reliability is calculated by finding the correlation between two or more raters as will be subsequently discussed.

III. Methodology

Being a Department of the University of Athens, the School of Foreign Languages adheres to the regulations imposed by the University. In particular, the School of Foreign Languages can only accept adult learners. Given that, it is assumed that foreign language learners are, in principle, perceptually, cognitively, linguistically and socially matured through their L1.

With respect to scoring procedures, a holistic rating is adopted based on a numerical scale ranging from 0-10. Its ultimate aim is to assess the overall proficiency level reflected in a given sample of student writing and it is generally considered a reliable scoring method, provided guidelines pertaining to rater training and rating administration are faithfully adhered to. As Perkins (1983) argues holistic scoring is the most reliable when the construct assessed is overall writing proficiency (p.652). Holistic scoring rubrics consist of 10 levels in the School of Foreign Languages

with marks 0-2 given to off topic essays, 3-4 to minimal evidence of proficiency, 5 to some evidence of proficiency, 6-7 to developing proficiency, 8 to proficient, 9 to very proficient and 10 to outstanding performance.

Along with the abovementioned scoring range, the School of Foreign Languages makes use of rubrics in order to analyze students' competency in certain features of language use and composition skills, such as content, organization, vocabulary, grammar, and mechanics. As far as content is concerned, attention is drawn to to the following aspects, namely, (a) the main idea is adequately developed, (b) relevant supporting details are provided and (c) there is sufficient knowledge of the subject matter. With respect to organization, the following parameters are taken into consideration: (a) there is no choppy or abrupt expression of ideas, (b) there is logical sequencing, (c) cohesiveness, (d) a clear and succint articulation of the main points and supporting details, and (e) ideas are not presented in a confusing way. Now turning to vocabulary, teachers are encouraged to look for: (a) correct and effective use of words in a given context, (b) advance use of vocabulary, and (c) appropriate use of register whereas when it comes to grammar, correct use of tense, number, word order, articles, pronouns, prepositions and complex constructions as well as absence of grammatical errors is highly valued. Lastly, mechanics refer to the adherence of spelling, punctuation and capitalization conventions, correct use of paragraphing and legible handwriting.

IV. Data Analysis and Discussion

Having outlined the rubrics on which L2 writing assessment is based on in the Greek School of Foreign Languages, I will now proceed with the calculation of the inter-rater reliability of the writing test in question. In particular, the following writing test was given to 12 adult students, 5 male and 7 female, of English proficiency level. Assessors were asked to put emphasis on the use - correct or incorrect - of the Future tenses since the below topic invites the use of such tenses.

Essay Topic

Will our lives be better or worse in 50 years from now?

(300-350 words).

For the purposes of the present study five randomly chosen essays amounting at least 40% of the total number of essays were selected as a representative sample as can be seen below:

Sample essay of the first student

I think that scientists will find cures for many diseases. Maybe the doctors will cure AIDS. With technology the food will be biological. So the food will be more health. Also, the quantity of food will be more and different. Researchers estimete that there will be another 2.5 bilion in the planet in fiftty years or so. Thus, we will need to provide food for them. It could be the case that new technologies goverments will have in order to prodoce foods. Moreover, the foods we eat will have a different format. Many people talk on artificial meat, which looks and feel like meat but it is not meat from an animal but from stem cells. We may even be eating artificial hambuger in a few years for all we know. To make things even more compicate, we may even have to experient eat insects, such as spiders, wasps, worms, ants, grashopers and beetles. Scientists will have invented environmentally – friendly cars, so the pollution levels will have decreased. The communication will be very simple through the internet. We can call America from Greece cheaper and faster. Also through the Internet will know our relationships. All these can make people lives longer.

However, the pessimistic people said that people life was worst in 50 years from now. They said that normal cities become too crowded so we will live in cities under the sea. The water there will not be more because we will destroy forests cycle. The human relationships will be gone because all day we will play in internet and will not read books. We will not want to meet other people. We will be afraid of meeting new people. We will not have friends and our life wil be miserable. We will work very hard in order to make ends meeting but our quality of our live will not necessarily be better. The relationships in our work eviroment will be hostile and very competiton. We wil not be ablle to trust our colleaggues because we will not believe that they are saying the tryth. As a result, we will fel lonely and isolate, and even sometimes secluded from the rest of the world.

Sample essay of the second student

I think people will be living longer because scientists will have found cures for many diseases like cancer. Also, children will be coming to the world healthy because scientists will modify the genes. Furthermore, I think in 50 years from now robbots will be doing homework. In my view, the way of education will be changed a lot and children will be studying through the internet, due to will lose human contact. School wil not have the format that has it in 50 years from now. There will be no more textbooks, paper, pens, or pencils and there wil be e-book only for students to read. There wil bi no ned to have teachers since lessons will be made through the internet. In this way, school will not exist but there wil bi e-clas for potential students. The education system will have advance significantly so as to bi able to help people of al age and learning abilitis to knowledges acquire in every filed of study. People will be able to do research and universities will open to everyone. Public education wil not be a dream byt a reality.

However, according to scientists, pollution levels in cities will increase and normal cities will become too crowed. As a result people will construct cities below sea. On the other hand climate will be changed all over the world and there won't be rainforests. Moreover, global warm wil have melted Antartica ice. In addition to that, people may be of danger becase of depletion of the ozone layer. Deaths from air pollution will be up and life of quality will be down. Unless we change the way we consume water and energy, there will be huge problem to face in the near future. There wil be a

greatest demand for other and extra resources because there will be two bilion people more in our planet in 50 years from now. How are we going to cope at such immense numbers? The answer to that qestion is not eiasier one to give nor a stretforward but I guess one first step that could make things better is people's awareness of the dangers that lie ahead of us.

Sample essay of the third student

I think that in fifty years from now, the image of our planet, will be quite different of how we know it. Trying to conceptualize Earth in 50 years from now, I imagine people living in cities under the sea because normal cities will have become too crowded. Also, I think that pollution levels in cities will have decreased since people will drive environmentally-friendly cars. Cars may even drive themselves in smart highwats. Companies will have developed tetchnology to such an extent so that car accidents will have being significantly reduced. Moreover, cars will be smaller in size but more effectively manipulated since they will have brains. In addition to that fuel economy will have increased as vehicles will be less thirsty for petrol. It could be the case that we may have the luxury to talk to our electrified car which will be be able to do much more than just provide us with simple driving instructions. For example, you may not need to learn how to park your car, because parking will have become an automated task inserted in the "brain" of your car. Furthermore, the quality of our lives will significantly improve because scientists will have found cures for many diseases.

However, no matter how optimistic somebody is, we cannot ignore some serious emerging problems as the years go by. The scary thing about all this technological advancement is that people may not even have the money to afford a car, let alone a high-tech car. Moreover, people will be too concerned with how many Facebook friends they have, instead of creating new friens and socializing in real-life contexts instead of internet ones. Values such as truth, friendship, love and freedom will have acquired a different meaning and nothing will be the same in the lives of young people who struggle to find their way in their society. Moreover, I cannot stop thinking that we will have burnt all the forests to construct more and more cities. Our food will have a different format and generally our planet will lose its natural habitat and will become more artificial. I believe that we must show our love to our planet as it is, before it is too late. Quality of life is not solely associated with material goods but there are other things, much simpler and down to earth, that can help us improve our lives. We are the ones tat decide what is important for us. A smile, a friendly handshake, a promise that is kept and a small favour that is made are only a few to mention.

Sample essay of the fourth student

I think that in 50 years from now many things will bi chaenge, as the example the way of living, will bi different. Nowdays cities will bi becoming too crowded, young people will leave from the small town and search for better live will go to big cities. The relationships also wil chaenge dramaticaly becauose the technologiy and the fast rhythms won't healp, they won't have free time for seinig and meeting new friends. Every kind of contact wil bi made through the Internet. Facebook will have a different format and there will be no need to meet someone outside. Every form of socializin wil bi made through Facebook. People wil bi isoleted ansd scered and there wil bi no quality in relationship because there wil not be any real-life relationship among people. We will become of our fellowmates even morre suspicious and wil be afraid to communicate to other people because we wil not know their intentions. As a result, we will spend more time alone watching T.V. and doing nothing productive.

However, I believe that scieintists fiend cures for many diseaseas, so people wil live longer. Also, scieintists invent environmental – friendly cars, thing that will help to keep the cities cleanest. Cars will have a brain. We will not have to drivi ours car because cars will be able to drive themselves. We will not have to learn how to pak our car because cars wil be automated and will do the parking by themselvs. Also, we will not need to have any fuel because cars will be electrical. There will be no car accidents because people will have become more skiled drivres. They will be more careful and concerned about oher people's lives. They will take a great care with pedestreans and pets as well.

Sample essay of the fifth student

I think that in 50 years from now, people will live in cities under the sea, as normal cities will have become too crowded. In addition to that, they will have invented environmentally-friendly cars, so pollution levels in cities will have decreased. More importantly, people will have a better quality of life because machines will do all the hard work since technology will have significantly improved. It may be the case that computers will outperform human beings. Of course, computers have a long way to go to match human strenghts but a considerable progress has been made in the processing power and memory capacity of the computers. Humanlike robots will have probably appeared, thus implementing a new way of life. Computers may even outperform human beings in most of their activities. Advancing computer performance will increase steadily and machines will begin to do well in various areas. In addition to that, I think that people will live longer in 50 years from now because scientists will have found cures for many diseases. In particular, scientists will be able to do experiments with human embryonic stem cells and find cures for many serious medical conditions such as cancer. Stem cell-based human therapies will have become a routine process and people will not be troubled by illnesses that seemed to have the cost the lives of many people in the past.

However, I also believe that there will be no more rainforests because we will have destroyed them all. Even though forests still cover 35 percent of the world's land area, the quality of land may have suffered such severe damage in the future so that there will be no forests. The most dramatic impact of this deforestation will be the loss of habitat for thousands of species. Another side-effect will be the climate change. Forest soils will quickly dry out if they are not

protected by trees and forest lands will be transformed into deserts. Furthermore, people will have become isolated since the only form of communication will be the internet. The end-result will be many socially-isolated people that are afraid to go out in the public and attemp social interactions. Consequently, they will become emotionally-isolated people that may have psychological dissorders leading to an unhappy and miserable way of life.

In spite of the above shortcomings, I tend to believe that our lives all in all will be better in 50 years from now for the simple reason that we will know how to handle good and bad things in a better way.

The aforementioned 5 written samples were assessed by five English language instructors, teaching in the School of Foreign Languages. All of them had a B.A. in English Language and Literature from the University of Athens and three of them had an M.A. in Applied Linguistics from a U.K. University whereas the other two had an M.A in TESOL again from a U.K. University. It should be mentioned that the raters were explicitly instructed not to signal mistakes in the essay papers to be marked. Hence, there were no identifying marks on the papers in order to ensure the maximum objectivity by leaving the raters completely uninfluenced. The results obtained are shown in Table 1 below:

TABLE 1:
ASSESSMENT OF FIVE WRITTEN SAMPLES

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Target Student 1	6	7	7	6	6
Target Student 2	7	8	7	7	6
Target Student 3	8	8	9	10	9
Target Student 4	5	5	6	6	5
Target Student 5	9	9	9	10	10

To have confidence in the ratings, we need information on inter-rater reliability. Since the students' final score is the combination or average of the ratings, reliability depends on the number of raters. According to Hatch and Lazaraton (1991) to compute inter-rater reliability for more than two raters, all the rating (producing a Pearson correlation matrix) must be firstly correlated and then an average of all the correlation coefficients must be derived (p.533). Given the above, to calculate the inter-rater reliability for a set of ratings made by 5 different raters we must firstly calculate ten correlations: the correlations between Rater 1 and Rater 2, Rater 1 and Rater 3, Rater 1 and Rater 4, Rater 1 and Rater 5, Rater 2 and Rater 3, Rater 2 and Rater 4, Rater 2 and Rater 5, Rater 3 and Rater 4, Rater 3 and Rater 5, Rater 4 and Rater 5. The general formula for calculating the correlation coefficient is the Pearson Product-moment correlation:

$$r_{xy} = \frac{\Sigma XY - (\Sigma X \Sigma Y)/n}{\sqrt{\left[\Sigma X^2 - (\Sigma X)^2/n\right]\left[\left[\Sigma Y^2 - (\Sigma Y)^2/n\right]\right.}}$$

where X represents the ratings of the first rater, Y represents the rating of the second rater, and n equals the number of target students rated by each rater (5 in our case). The symbol Σ indicates that you need to sum the relevant values. Thus, ΣX indicates that you need to sum the values of X, and ΣXY indicates that you need to sum the values of X multiplied by Y (Maclennan, 1993). Hence, by substituting these specific ratings into the general formula for a correlation, we come up with the following Pearson correlation matrix:

TABLE 2:
PEARSON CORRELATION FOR RATERS

R1	**R2**	**R3**	**R4**	**R5**
R1	.93	.94	.52	.94
R2		.83	.42	.80
R3			.54	.97
R4				.54
R5				

Now, using these correlations, we can calculate inter-rater reliability using the Spearman-Brown formula:

$$r_a = \frac{nr_{AB}}{1+(n-1)^r AB}$$

where r_{ab} is equal to the average correlation between the ratings made by each rater. The average for these correlations would be:

$$r_{ab} = \frac{7.43}{10}$$

$$r_{ab} = .743$$

The inter-rater reliability is:

$$r_a = \frac{(5)(.743)}{1+[(5-1)(.743)]}$$

$$r_a = \frac{3.715}{1+(4)(.743)}$$

$$r_a = \frac{3.715}{1+2.972}$$

$$r_a = \frac{3.715}{3.972}$$

$$r_a = .93$$

Thus, the overall inter-rater agreement for the five markers is **.93**. According to Landis and Koch (1977) values less than 0.00 indicate poor agreement, values between 0.00 and 0.20 indicate slight agreement, values between 0.21 and 0.40 indicate fair agreement, values between 0.41 and 0.60 indicate moderate agreement, values between 0.61 and 0.80 indicate substantial agreement, and values between 0.81 and 1.00 indicate almost-perfect agreement. Thus, a value of 0.80 and above is almost perfect. Consequently, this is a very satisfactory value which leads us to infer that the reliability is very high as well as the level of rater agreement. The main factors that have contributed to the test's reliability are first and foremost, the fact that the aim of the writing assessment was straightforward, namely, judges were asked to assess students' use of Future tenses, and secondly, there was a clear and economical scoring scale as well as rubrics to safely guide assessors in their marking of the essays in question. Thus, clarity of instructions and a clear, unambiguous marking scheme are essential steps in ironing out differences in teacher interpretation of the assessment criteria applied to a given test.

V. Conclusion

In this paper, an attempt was made to calculate the inter-rater reliability of 5 written samples of proficiency-level students attending the Greek School of Foreign Languages. Results have shown a very satisfactory level of inter-rater agreement between the five judges. Given these results, it can be concluded, than on a broader research front, similar, longer-scale studies can offer useful insights into the writing construct, with the identification and development of criteria which are both useful and relevant for the assessment of both L1 and L2 writing.

References

[1] Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.
[2] Fulcher, G. (2010). Practical Language Testing. London: Hodder Education.
[3] Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment.* London and New York: Routledge.
[4] Fulcher, G. (1999). Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics* 20 (2), 221-236.
[5] Fulcher, G. (1997). Assessing writing. In Fulcher, G. (ed.) *Writing in the English Language Classroom* (pp.91-107). Prentice Hall Europe in association with the British Council. ELT Review Series.
[6] Gamaroff, R. (2000). Rater reliability in language assessment: the bug of all bears. *System* 28, 31-53.
[7] Hamp-Lyons, L. (1990). Second language writing: assessment issues. In Kroll, B. (ed.) *Second Language Writing* (pp.69-87). Cambridge: Cambridge University Press.
[8] Jones, R. (1979). Performance testing of second language proficiency. In Briere, E. and Hinofotis, F. (eds.) *Concepts in Language Testing* (pp. 50-57). Washington, DC: TESOL.
[9] Hatch, E. and Lazaraton, A. (1991). The Research Manual: Design and Statistics for Applied Linguistics. Boston, Massachusetts: Heinle & Heinle Publishers.
[10] Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33 (1), 159-174.
[11] Lombard, M., Snyder-Duch, J. and Campanella-Bracken, C. (2004) Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved on September 23, 2012 from: http://www.temple.edu/ispr/mmc/reliability/#What%20is%20intercoder%20reliability.
[12] Maclennan, R.N. (1993). Inter-rater reliability with SPSS for windows 5.0. *The American Statistician* 47 (4), 292-296.
[13] McNamara, T. (1996). Measuring Second Language Performance. London: Longman.
[14] O'Sallivan, B. (2011). Language Testing: Theories and Practices. Basingstoke: Palgrave Macmillan.
[15] Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly* 17, 651-671.
[16] Tinsley, H.E.A. and Weiss, D.J. (2000). Interrater reliability and agreement. In Tinsley, H.E.A. and Brown, S. D. (eds.) *Handbook of Applied Multivariate Statistics and Mathematical Modeling.* (pp. 95-124). San Diego, CA: Academic Press.

Despoina Panou is a Ph.D. student at the University of Lecester. She has a B.A. in English Language and Literature (with distinction) and an M.A. in Translation-Translatology (with distinction) from the University of Athens. She also has an M.A. in Linguistics-TESOL (with merit) from the University of Surrey. Her research interests are foreign language teaching, translation, and the use of figurative/idiomatic language in business discourse.