Inter-rater reliability in the Greek State Certificate of Language Proficiency (KPG) exams

Vassilis Hartzoulakis

Research Centre for English Language Teaching, Testing and Assessment
University of Athens

Abstract

This paper investigates the extent to which KPG script evaluators agree on their scoring decisions. The overall aim of the study is to check the effectiveness of the instruments subserving the rating process as designed by the KPG experts. Data for the B2 and C1 levels written production tasks for all the KPG exam periods in the years 2005, 2006 and 2007 for the B1 level for the year 2007 were gathered and analysed in terms of reliability in scoring decisions. The analysis of the data shows that raters demonstrated quite high correlations in the corresponding examination periods. The inter-rater reliability index is constantly kept above r=.50 which qualifies them as strong correlations. Consistently higher correlations for the ratings in level B2 than in level C1 were also recorded. Overall, the analysis of the data shows that raters apply the criteria set in the rating grid in a uniform way, demonstrating strong significant correlations in the different rating periods.

Keywords: testing, script rating, validity, inter-rater reliability, rater agreement, correlation

1. Introduction

This paper investigates rater agreement in the 'State Certificate of Language Proficiency' (KPG) module 2 (free written production and written mediation) exams. We look into the factors affecting the rating process in the specific situation and examine the extent to which these factors contribute to higher inter-rater reliability. The paper presents an analysis of the rater agreement for the examinations in the B2 and C1 levels in the years 2005, 2006 and 2007 and for the B1 level examinations in the May and November 2007 periods. For the purposes of the present study a number of randomly selected papers amounting to at least 40% of the total number for each level and each period were selected as a representative sample. The inter-rater reliability index was then computed separately for each of the two tasks that comprise the whole of the written part of the exam.

2. Aims of the study

The on-going study on inter-rater reliability in the KPG script rating process is carried out as a means of investigating the effectiveness of the instruments subservient to the process employed by the KPG test developers. These instruments are: (a) the rating grid together with the assessment criteria, (b) the script raters training material and training seminars and (c) the on-the-spot consultancy to the raters by KPG experts and test developers. The training material and training seminars are individualised for every single period based on the given tasks, resulting in specific instructions as to how each different writing task should be rated. The same applies for the consultancy provided to the script raters during the actual process of rating, which adds to the homogeneity of the rating grid interpretation. The abovementioned instruments have been designed with the aim of achieving the highest possible rater agreement, which is part and parcel of the overall reliability for any test.

3. Approaches to rating scripts

Expressing thoughts in written form is 'probably the most complex constructive act that most human beings are ever expected to perform' (Bereiter and Scardemalia 1983: 20 cited in Gamaroff 2000). The complexity of the act makes the objective assessment of performance very difficult. The way a reader/script rater understands a written text -especially in essays or even short compositions where inferential judgements have to be made- varies depending on factors that have to do with the individual's global comprehension of a passage, his or her inferential ability, and his or her interpretation of meaning of words in each context. Hamp-Lyons (1990) argues that the reliability of rating scripts heavily depends on the attitudes and conceptions of the rater. Therefore, problems arise in evaluating objectively when inferential judgements have to be converted to a score. It is true that one can have a largely objective scoring system when scores are primarily based on correct structural forms, as is the case with numerous language exams. However, this is not applicable in the KPG system as it does not focus on measuring correct syntactico-grammatical forms only, but on measuring "candidates' abilities and skills to make socially purposeful use of the target language at home and abroad" (RCEL 2007). This does not simply entail correct syntax and grammar in the produced written texts, but also making appropriate language choices by taking into consideration the communicative and social context within which the produced text appears.

The question that comes up then is: How can writing tasks be converted to numbers that will yield meaningful variance between learners? It has been suggested (Oller 1979) that inferential judgements should be based on intended meaning and not merely on correct structural forms. Gamaroff (2000) suggests that when rating written texts, it is preferable for the rater to 'rewrite' the intended meaning in his or her mind and then decide on a mark. Still, even then, one cannot secure reliability and objectivity in rating as there are different conceptions of appropriately conveyed meaning or not.

Another approach to rating written texts is correcting grammatico-syntactical and lexical mistakes and accordingly subtracting points for every one depending on its seriousness. This approach has been heavily criticised as each rater has his or her own standards regarding what is grammatically correct or not, let alone the concept of the seriousness of mistakes, which also varies in the mind of every rater.

The above led researchers to the construction of rating grids with the purpose of aiding the rater when converting the text's qualitative characteristics into quantitative ones. Rating grids have been found to contribute to the decrease in subjectivity when it comes to rating written texts, although they cannot secure absolute objectivity (Τσοπάνογλου 2000). Evidently rating grids should be explicit and concise. They should be explicit so that raters will interpret them homogeneously; and they should be concise so that they are practical to use. When rating grids are properly employed by trained raters, the rating procedure will most likely display consistency among raters.

4. KPG writing, script rater training and the rating procedure

The KPG examination system is organised in four modules as follows:

1. Module 1 which tests reading comprehension and language awareness,
2. Module 2 which tests free written production and written mediation,
3. Module 3 which tests listening comprehension and
4. Module 4 which tests free oral production and oral mediation.

Module 2 requires candidates to produce two texts of various lengths that range from 100 to 300 words each, depending on the exam level. The first text is produced based on stimulus given in the target language (English in our case, although the examination systems provides batteries for French, German, Italian and Spanish as well) whereas the second one is produced based on stimulus given in Greek. In this case candidates have to act as mediators selecting information from the Greek source and transferring it to English either in similar or even completely different formats. Each script is rated twice

by two different script raters. The script raters rate each of the two tasks on a scale 0-15 employing the rating grid and assessment criteria set by the test developers without signalling anything on the papers themselves, then mask their marks and names and return the papers to the examination secretariat. The rated papers are then randomly redistributed to the same pool of raters, taking care that no paper is given to the same rater that initially marked it.

4.1 Training KPG script raters

Before the rating procedure for every examination period begins, the KPG English team prepare a training seminar for all script raters where the candidates' performance expectations for the specific test are presented and discussed. The performance expectations are individualised for every single test and are determined beforehand in the piloting phase of the test before its administration and in pilot-evaluating sessions held after its administration. During the training seminar, raters have the chance to go through the rating grid (Appendix) in conjunction with the performance expectations and rate sample papers by applying the criteria that have been set for every different task. This ensures the adoption of a common approach towards rating and helps in establishing consistency in the given marks. This kind of training is an on-going process, as during the rating procedure itself, each script rater is assigned to a supervisor (an experienced and specially trained member of the KPG personnel) who constantly offers support by discussing fine points and offering his or her opinion in cases where the rater is uncertain about the proper employment of the criteria for assessing the paper.

4.2 Rating procedure

The script raters rate the two texts produced by each candidate on a scale of 0 to 15 for each text and on the basis of the rating grid discussed previously. Candidates' papers are grouped in packs of 50 and are rated by two raters randomly selected from a pool of about 150. After the 1st rater has rated the 50 papers in a pack and the given marks are masked, the pack is passed to a 2nd rater who also gives his or her own marks. The final mark given to each candidate is the mean score between the two raters. It is interesting to note that contrary to prevalent practices in other high stakes examination systems, the raters do not signal mistakes (either stylistic or structural) on the papers; therefore the 2nd rater does not see any notes made on the paper by the 1st rater and is left completely

uninfluenced, which ensures the maximum possible objectivity. Bachman (2004) argues that it is essential that the second ratings be independent of the first ones and if written essays are scored, no identifying information and no marks should be made on the paper during the first rating. On the other hand, this procedure runs a bigger risk of inconsistencies in the way the two raters rate the responses resulting in measurement errors; still, this issue is resolved by training the raters as meticulously as possible on how to employ the rating grid in combination with the candidates' expected outcomes for every single exam, then fine tune and re-evaluate the procedure by constantly estimating the consistency across raters, or in other words, the inter-rater reliability of scores.

5. Approaches to inter-rater reliability

Inter-rater reliability is the widely used term for the extent to which independent raters evaluate data and reach the same conclusion (Lombard et al. 2005). It is part of the overall analysis for rater agreement, which is concerned with reconciling the raters' subjectivity and the objective precision of the mark. Inter-rater reliability investigation is particularly important in 'subjective' tests such as essay tests, where there exist fluctuations in judgements between different raters (Gamaroff 2000). However, agreement among raters is extremely important not only in academic domains but in every domain where more than one judge rates performances. Such domains include areas as far apart from each other as gymnastics or figure skating in the Olympic Games, medical diagnoses, jurors' judgements in criminal trials and test-eaters' judgements on the chef's performance when rating restaurants (Von Eye and Mun 2005).

Inter-rater reliability studies in education mostly focus on the consistency of given marks to establish the extent of consensus on use of the instrument (rating grid) by those who administer it. In such cases, it is vital that all raters apply the criteria on the rating grid in exactly the same way, resulting in a homogeneous rating approach. This, in turn, is one of the criteria that comprise a reliable testing system. Tinsley and Weiss (2000) prefer the term 'inter-rater (or inter-coder) agreement' as they note that although inter-rater reliability assesses how far "ratings of different judges are the same when expressed as deviations from their means," inter-rater agreement is needed because it measures "the extent to which the different judges tend to assign exactly the same rating

to each object" (p. 98); However, here, the term inter-rater reliability will be used in its widely accepted sense, as a correlation between the two sets of ratings (Bachman 2004).

Statisticians have not reached a consensus on one universally accepted index of inter-rater reliability and depending on the type of data and the purpose of the study, different indices have been suggested. Some options are: joint-probability of agreement when the rating scale is nominal, Cohen's kappa and the related Fleiss' kappa for two raters and a nominal or ordinal scale, inter-rater correlation, concordance correlation coefficient and intra-class correlation. When measuring correlation among pairs of raters using a scale that is ordered perhaps the most popular statistic is the Pearson correlation coefficient or 'Pearson's r' (Stemler 2004). It is a convenient index as it can be computed by hand or by using most statistical software packages. If the rating scale is continuous, Pearson's r can be used to measure the correlation among pairs of raters. If the rating scale is ordinal, Spearman's ρ is used instead. However, in both cases, the magnitude of the differences between raters is not taken into account. Shrout and Fleiss (1979) demonstrate this drawback with an example: If Judge A assigned the scores 9, 10, 5, 15 to four scripts and Judge B assigned 7, 8, 3, 13 to the same scripts (the difference is consistently kept at -2 points for all four scripts), then using Pearson's method, the correlation coefficient would be 1,00, indicating perfect correlation, which is definitely not the case in this example. Instead of Pearson's r, Shrout and Fleiss (1979) suggest calculating the intra class correlation coefficient (ICC) as another way of performing reliability testing. The ICC is an improvement over Pearson's as it takes into account the differences in ratings for individual segments, along with the correlation between raters. In the example above, the ICC is .94, a measurement which is more representative of the case. All in all, the ICC should be used to measure the inter-rater reliability for two or more raters and especially if we are interested in using a team of raters and we want to establish that they yield consistent results.

The KPG system employs several judges who randomly form pairs. This, together with the fact that there are no signals or notes on the candidate's paper after the first rating, leads to handling the raters as absolutely equal variables. That is, within any pair of ratings, there is no reason to identify one as 'first' and the other 'second'; if some or all of them are labelled the other way round the calculated correlation presumably would not change. According to Shrout and Fleiss (1979) there are numerous versions of the ICC that can give quite different results when applied to the same data. Therefore, careful consideration of the data layout and the ICC version is required if one is to come

up with a valid index. When computing the ICC, the data should be laid out as N cases or rows, (each row corresponds to each script) and k variables or columns, for the different measurements (first and second rating) of the cases (Wuensch 2007); in our model there are two different measurements/ratings for every script. The cases are assumed to be a random sample from a larger population, and the ICC estimates are based on mean squares obtained by applying analysis of variance (ANOVA) models to these data. ICC varies depending on whether the judges in the study are all the judges of interest or are a random sample of possible judges, whether all targets are rated or only a random sample, and whether reliability is to be measured based on individual ratings or mean ratings of all judges (Shrout and Fleiss 1979). When the judges/raters are conceived as being a random selection of possible raters/judges, then a one-way ANOVA is employed. That is, in this model judges are treated as a random sample and the focus of interest is a one-way ANOVA testing if there is a subject/target effect (Garson 1998).

Intra class correlations in general, are considered to be measures of reliability or measures of the magnitude of an effect, but they are equally important when it comes to calculating the correlations between pairs of observations that don't have an obvious order (Maclennan 1993). The intra class correlation coefficient can be easily computed in SPSS and other statistics software packages. There are five possible sets of output in the ICC estimates as offered in the SPSS; of those, the one most appropriate for computing the ICC in the KPG examination system is the one-way random effects model with an estimate for the reliability for the mean for average measures. In this model, judges/raters are conceived as being a random selection of possible raters/judges, who rate all targets of interest. Even though in this study not all targets of interest are measured, we still need to select the one-way random effects model because this model applies even when a given rating (ex., the first rating) for one subject might be by a different judge than the first rating for another subject, etc. This in turn means there is no way to separate out a judge/rater effect (Garson, 1998).

The ICC can take any value between 0.00 (which signifies no correlation) and 1.00 (when there is no variance within targets). Statisticians give different interpretations of ICC values, but two of the most widely accepted interpretations are those of Fleiss (1981) and Landis and Koch (1977) presented in Tables 1 and 2, respectively.

Table 1. ICC interpretation according to Fleiss (1981)

r <0.40	poor agreement
0.40≤ r ≤0.75	good agreement
r >0.75	excellent agreement

Table 2. ICC interpretation according to Landis and Koch (1977)

r <0.00	poor agreement
0.00 ≤r ≤0.20	slight
0.21 ≤r ≤0.40	fair
0.41 ≤r ≤0.59	moderate
0.60 ≤r ≤0.79	substantial
0.80 ≤r ≤1.00	almost perfect

Based on the information in the tables above, we can assume that a value of 0.60 and above can be considered to represent a satisfactory intra class correlation, implying a satisfactory level of rater agreement.

6. Findings

Data from the examination periods in the years 2005-2007 were gathered and analysed using SPSS. All correlations are statistically significant ($p<.05$) and when checked for the Tukey's test of non-additivity, showed that there is no multiplicative interaction between the cases and the items.

Table 3. ICC measurements for all levels and periods

	MAY 2005	NOVEMBER 2005	MAY 2006	NOVEMBER 2006	MAY 2007	NOVEMBER 2007
Free Writing Production						
B1					.76	.73
B2	.74	.70	.76	.68	.76	.72
C1	.57	.56	.63	.52	.59	.66
Mediation						
B1					.83	.88
B2	.77	.75	.74	.72	.80	.69
C1	.62	.60	.68	.53	.69	.71

Table 3 above shows that with the exception of C1 level in the November 2006 period, the ICC for both tasks (free writing production and mediation), for all levels and periods is either well above or slightly below the cut-off score of 0.60 that we set as representative of a satisfactory agreement. The ICC for B1 level is slightly higher than

that for the B2 level, which in turn is higher than that for the C1 level. This is also reflected in the Figures 1 and 2 below which show the ICC fluctuation for the free writing production and mediation, respectively.

Figure 1: ICC Free Writing Production Timeline for all levels

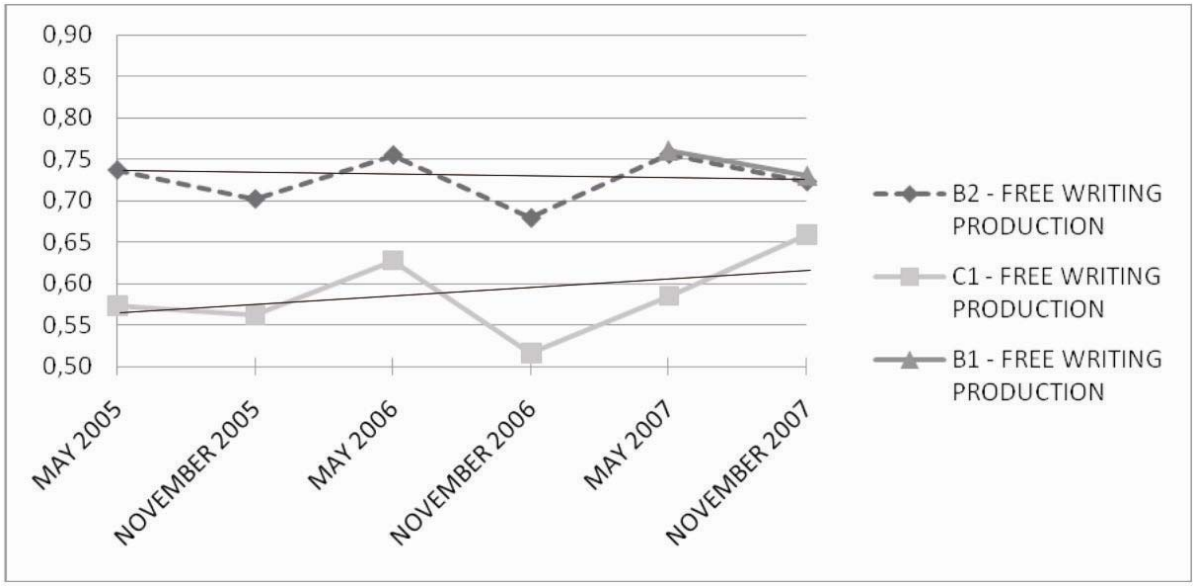
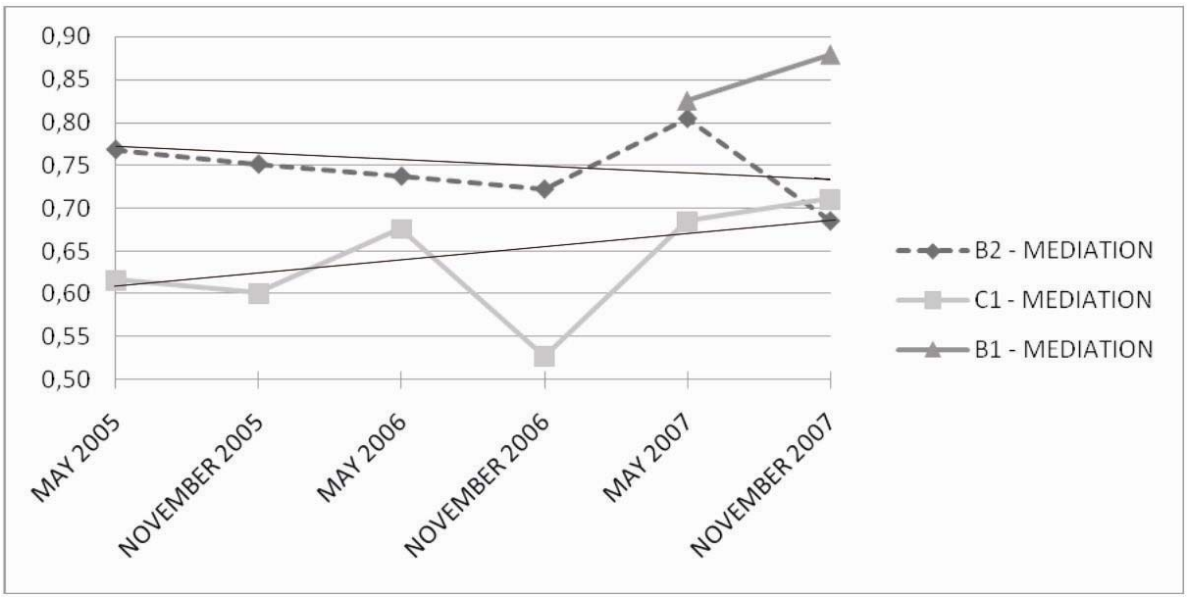


Figure 2: ICC Mediation Timeline for all levels



Figures 1 and 2 also clearly demonstrate the established pattern of lower ICC as the test level becomes higher. A closer look at figure 1 shows a tendency for a more or less stabilized ICC index for the B2 level at around .70, whereas the ICC index for the C1 level shows an upward-sloping trendline converging with the B2 measurements. The

same pattern is followed in the data presented in figure 2, with the two lines converging around the .70 measurement. Even though data for the B1 level are not yet sufficient for any definite conclusion, one sees that ICC estimates for the free writing production (Figure 1) are slightly higher than those of B2 level and seem to be converging towards a little above .70. The B1 indices for mediation (Figure 2) are significantly higher, moving well above .80, but this remains to be verified with subsequent measurements.

7. Discussion of findings

The analysis of the data obtained for the two examination periods of the KPG system for the years 2005, 2006, and 2007 shows that raters demonstrated quite high correlations in the corresponding examination periods. The ICC is for most of the cases kept above r=.60 which qualifies them as strong correlations (Cohen 1988). One can also notice that the ICC estimates are consistently higher for the ratings in level B2 than in level C1. This can be attributed to the fact that raters are more experienced in rating B2 level papers, as this specific exam was the first to have been administered by the Hellenic Ministry of Education and Religious Affairs, almost a year and a half before the C1 level exam was introduced. Therefore, there were two rating periods where raters rated only papers at B2 level, before they started rating papers at C1 level. Additionally, we can assume that C1 level scripts demonstrate more complex language choices and deeper and broader cognitive processes, which leave raters with a broader range of decisions.

The fact that the ICC for B1 level (which is the latest addition in the system) is higher than that of B2 level (although it is still early to establish a fixed pattern) can be attributed to various factors. One can be the lower language level for that test which makes script rating simpler in terms of linguistic and syntactical choices and judgements. A second factor is that the candidates (and consequently the script raters) are given a sample script which they have to follow in the actual test. This script is acting as a guide for the candidates when producing a script resulting in a more or less homogenous approach to the requested task.

There is a slight drop in ICC estimates in the November 2006 examination period. As one clearly sees in figures 1 and 2, this drop is reflected in both B2 and C1 levels and in both tasks. Since in that examination period quite a large number of new raters were introduced into the system, one might assume that the experience in rating KPG scripts factor was affected, resulting in this drop in ICC. The effect is rectified in the following

periods, which leads us to the assumption that experience in rating is of the utmost importance when it comes to rater agreement. It is worth noting that there were very similar correlations between the free written production and mediation tasks for each period and for each level when examined individually. This implies that raters exercise a uniform approach towards applying the criteria set for every period and every separate exam. If one looks at correlations throughout the two levels in the last three years, one cannot fail to see that there are no extreme fluctuations in the strengths. This is another indication of uniformity in the overall approach towards rating written texts in the KPG system.

8. Conclusion

The analysis of the data yielded from the KPG script rating shows that raters apply the relevant criteria in a generally uniform way, showing strong significant correlations in the different rating periods. The tools employed by the test developers have a positive effect on rater agreement indices as ICC estimates follow an upward-sloping trendline. This implies that experience in rating leads to better correlations, thus constant training of the raters on the part of the test developers is required.

References

Bachman L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Cohen J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Fleiss J.L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley and Sons, Inc.

Gamaroff R. (2000). Rater reliability in language assessment: The bug of all bears. *System* 28: 31-53.

Garson D.G. (1998, 2008). Reliability analysis. http://www2.chass.ncsu.edu/garson/pa765/reliab.htm

Hamp-Lyons L. (1990). Second language writing: Assessment issues. In B. Kroll (ed), *Second language writing*. Cambridge: Cambridge University Press, 69-87.

Landis J.R. and G.G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33/1: 159-174.

Lombard M., J. Snyder-Duch and C. Campanella-Bracken (2005, June 13). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. http://www.temple.edu/ispr/mmc/reliability/#What%20is%20intercoder%20reliability

Maclennan R.N. (1993). Inter-rater reliability with SPSS for windows 5.0. *The American Statistician* 47/4: 292-296.

Oller J.W. (1979). *Language tests at school*. London: Longman.

RCEL (2007). The Greek State Certificate of Language Proficiency. http://www.cc.uoa.gr/english/rcel/texts/KPGdescription-2008.pdf

Shrout P. and J.L. Fleiss (1979). Intra class correlation: Uses in assessing rater reliability. *Psychological Bulletin* 86/2: 420-428.

Stemler S.E. (2004). A comparison of consensus, consistency, and measurement: Approaches to estimating inter-rater reliability [Electronic version]. *Practical Assessment, Research and Evaluation,* 9/4. http://PAREonline.net/getvn.asp?v=9&n=4

Tinsley H.E. and D.J. Weiss (2000). Inter-rater reliability and agreement. In H.E.A. Tinsley and S.D. Brown (eds), *Handbook of applied multivariate statistics and mathematical modeling*. San Diego, CA: Academic Press, 95-124.

Τσοπάνογλου Α. (2000). *Μεθοδολογία της επιστημονικής έρευνας και εφαρμογές της στην αξιολόγηση της γλωσσικής κατάρτισης*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Von Eye A. and E.Y. Mun (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Lawrence Erlbaum.

Wuensch K.L. (2007). Inter-rater agreement. http://core.ecu.edu/psyc/wuenschk/docs30/Inter-rater.doc

Appendix

Assessment criteria for rating module 2 scripts

1. Text content, form, style and organization	**2. Appropriacy of lexicogrammatic selections**	**3. Appropriacy of linguistic expression cohesion and coherence of discourse**	**G R A D E**
Production of a written text of specific *content**, in accordance with the communication case as defined in the directions, which describes the choice of *form*, *styleç* and *organization* of the form of speech prescribed by the "norm" (eg, advertisement, application, report).	Selection of the proper linguistic elements (words and phrases), given their textual and contextual framework.	Appropriate grammatical and syntactical use of language with coherence and cohesion.	
Very satisfactory	A text that responds successfully to the 3 criteria.	Development of general meaning and partial meanings with acceptable uses of language and cohesion of speech.	**15**
		It has got minimum errors which do not prevent the transfer of the meaning of the text	**14**
	A text that refers to the subject of the test and has got the required form and organization. It includes some mistakes which do not essentially obstruct communication. In general, it is a text with a natural flow of discourse.	It includes generally accepted uses of language, although the linguistic selections may not always be the most appropriate.	**13**
		Certain linguistic selections are not appropriate but in accordance with the basic grammar rules.	**12**
Moderately satisfactory	A text that does not deviate from the subject of the test and is in the demanded form. It includes some errors that hinder the transfer of meaning. In general, it is a text of not totally natural flow of speech and cohesion of phrases.	Certain linguistic selections are not appropriate and in some cases deviate from rules of language usage, but the diction is satisfactory.	**11**
		Several linguistic selections are inappropriate, diction relatively satisfactory but some phrases are awkward.	**10**
	A text that deals with aspects of the subject and approaches the form requested by the task. It includes errors that hinder	Certain linguistic selections are not appropriate and deviate from the acceptable use of language.	**9**

	understanding at some points, but there is relevant cohesion of discourse.	Several linguistic selections are inappropriate and/or incorrect according to grammar rules.	**8**
Partly satisfactory	A text that does not deal with the subject in an absolutely satisfying a manner and does not exactly have the form requested by the task. To an extent errors inhibit its general understanding.	Limited vocabulary, inappropriate expressions, errors but the meaning is transferred.	**7**
		The general meaning is transferred but the particular information is difficult to understand.	**6**
	A test which does not have the form requested by the task and includes errors of various types.	Many errors significantly hindering the understanding of main points.	**5**
		Many and serious errors of vocabulary, grammar, spelling, etc.	**4**
Not satisfactory * For Activity 2: based on a prompt in Greek	Irrelevant		**3**
	Text not understood		**2**
	Words scattered		**1**
	No answer		**0**