Future Academy®
Offering excellence in Social and Behavioural Sciences

The European Proceedings of
Social & Behavioural Sciences
EpSBS

eISSN: 2357-1330

**WLC 2016 :** World LUMEN Congress. Logos Universality Mentality Education Novelty 2016 | LUMEN 15th Anniversary Edition

# The Role of Monitoring Raters in Ensuring Accurate and Meaningful Test Scores. Case Study: RFL Examinations

Lavinia Vasiu[a]*, Antonela Arieșan[b]

* Corresponding author: *Lavinia Vasiu, laviniavasiu@gmail.com*

[a]*Babeș-Bolyai University, Cluj-Napoca, Romania, laviniavasiu@gmail.com*
[b]*Babeș-Bolyai University, Cluj-Napoca, Romania, antonela.suciu@gmail.com*

**Abstract**

*http://dx.doi.org/10.15405/epsbs.2016.09.132*

Providing accurate and meaningful test scores is an extremely important issue especially in the case of high-stakes examinations like the one considered here: the RFL examination, level B2, which practically conditions the admission of students to an academic program in Romanian. The paper aims to describe and explain how data regarding the achievement of raters are collected and analysed in order to ensure rating accuracy and rater reliability. Monitoring, co-ordination, standardization measures all aim at dealing with problems of leniency, inconsistency or severity of raters. The paper details the procedures used for calculating rating accuracy, intra-rater reliability, inter-rater agreement in the case of marking both receptive and productive components of the RFL examination.



**Keywords:** Raters; monitoring; standardization; reliability; accuracy; agreement.

## 1. Introduction

The examination taken into consideration here is the test of Romanian as a foreign language (RFL), level B2. This can be regarded as a high-stakes test as it represents the test one has to pass in order to obtain a linguistic competence certificate in Romanian which usually conditions the admission in any academic program taught in Romanian in any university in our country. Currently, the two categories of population taking this test are: the students enrolled in Babeș-Bolyai University, in the preparatory year (Faculty of Letters, Department of Romanian culture, language and civilization), and the persons who simply need the certificate in Romanian in order to be able to start studies in Romania (60-130 candidates per year).

Being a high-stakes test, it is only natural that the organization providing it should regard as important all aspects concerning its quality. Therefore, the Department of Romanian culture, language and civilization submitted the test to be audited by ALTE (Association of Language Testers in Europe) and in 2015 obtained the ALTE Q-mark (a quality indicator showing that the exams provided by the organization "have passed a rigorous audit and meet all 17 of ALTE's quality standards" and allowing test users "to be confident that an exam is backed up by appropriate processes, criteria and standards" (www.alte.org). The validity argument presented evidence of validity for all aspects of the assessment process: test development, item writing, test administration, marking and grading, reporting of results, etc. (Hughes, 1989).

Validity in testing and assessment is defined as discovering whether a test "measures accurately what it is intended to measure" (Messik, 1989, p. 22). Messick saw validity not as a property of a test or assessment, but as the extent to which one is allowed to make inferences to a construct from a test score and the degree to which any decision one might make on the basis of the score is justifiable (AERA, APA & NCME, 1985, p. 13). This definition of validity has become the accepted paradigm in psychological, educational and language testing: "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the score. The inferences regarding specific uses of a test are validated, not the test itself." (Fulcher & Davidson, 2007, p. 9).

In what concerns language testing, as we have mentioned above, part of a validity claim is that the test administration and all the processes used by the testing agency are done according to standardized procedures and one of the most important aspects of monitoring the quality of those is identifying the key stages, describing what needs to be done and to what standards and comparing what is actually done to these standards (Green, 1998, p. 127-128).

The present study focuses on one source of evidence for validity: the grading process. Accuracy of rating and reliability of test scores are analysed by monitoring raters' activity.

http://dx.doi.org/10.15405/epsbs.2016.09.132
eISSN: 2357-1330 / Corresponding Author: Lavinia Vasiu
Selection and peer-review under responsibility of the Organizing Committee of the conference

## 2. Participants

The participants in the study were 10 professional raters involved in the examination process during the academic year 2015 (the B2 exam, the Department of Romanian language, culture and civilization, Faculty of Letters, UBB). The staff involved in the assessment process is carefully selected. The requirements for markers include: BA or MA in Romanian language, an academic degree in teaching and assessing Romanian as a foreign language (teacher training module/MA in teaching RFL), at least 2 years of experience in teaching RFL and in preparing students for RFL examinations, familiarization with the procedures to follow, with the mark scheme and with the answer key (with special focus on partial credit items, for listening comprehension and for elements of communication construction), attendance of group training sessions (standardization and training for assessment, in the case of productive skills).

Each rater assessed a number between 9 and 12 students/papers. Thus, 106 papers assessed by 10 raters were taken into consideration for the study.

## 3. Methodology outline

In the study both qualitative and quantitative methods have been used. In the case of the RFL examinations (which have 5 components: Listening, Reading, Elements of Communication Construction - ECC, Writing and Speaking), the first three components are formed from questions which allow objective marking. These components are double-marked in order to ensure that the mistakes are avoided/repaired. After the first marker corrects the three components and inserts the results in the corresponding boxes, on the second page of the exam paper, the second marker performs the second correction. If he/she arrives at a different score for any of the three components, he/she checks his/her own correction as well as the first marker's in order to discover possible errors of calculation or mistakes in the application of the mark scheme. If such an error is discovered, then the marker makes the correction and registers the change. During the marking period, markers can ask for feedback (as the rating task is performed in the same room, under the supervision of the chief examiner) but they are also randomly checked by the chief examiner as they carry out their task.

The mark scheme is consistent from one session to the other and the answer key is unambiguous. For the item types used (multiple choice, true/false, matching, gap-filling) there is either only one possible correct answer or a very limited and clearly defined number of acceptable answers (partial credit items). As mentioned above, the assessors are aware of the procedures to follow, they are familiar with the mark scheme and they get familiarized with the answer key before each session of examination. As the questions in these three components allow objective marking, in these cases, raters were only checked for accuracy (how well they apply the marking scheme). Data regarding the number and the types of errors each rater makes was gathered and processed, as shown in Table 1:

**Table 1**. Data regarding the types of errors – example

| Rater 1 | Rater 2 | Rater 3 |
|---|---|---|
| 3 L + 1 C + 8 ECC = 12 | 2 C + 2 ECC = 4 | 1 C + 2 ECC = 3 |

ECC = error when marking the elements of communication construction component
L = errors when marking the listening component
C = errors of calculation

$$\text{Error rate} = \frac{\text{total number of errors/rater}}{\text{total number of papers rated by a rater} * \text{total number of items/paper}}$$

**Table 2 .**Error rate – example

| Rater 1 | Rater 2 | Rater 3 |
|---|---|---|
| 1/9x90 = **0.00123** | 9/9x90 = **0.01111** | 1/10x90 = **0.00111** |

As mentioned before, between 9 and 12 papers were considered in the case of each assessor. If the error rate was too high (higher than 0.02), the assessor will be carefully observed during the following sessions of examination.

Also, at this level, types of errors that occur within raters were identified in order for the board of examination to prepare the following training sessions for raters focusing on the most significant aspects that need to be improved (e.g. partial credit items).

In what concerns the Writing component, each script is analytically marked by two examiners who use a detailed grid and assessment forms on which they write their comments justifying the score given for each criterion. If they give widely varying marks (a difference of more than 2 points in the final marks), the script is marked a third time by the chief examiner. The number of points he/she grants is taken into consideration as well when calculating the average which represents the final score. If no such situation appears, the overall score for the writing component is represented by the average of the two marks given by the two assessors. One method of monitoring raters adopted in the case of RFL examination is the use of pre-assessed scripts (Council of Europe, 2001, p. 43). The chief examiner performs this task himself. He makes copies of the written productions of some of the candidates and rates them, placing the papers back in the pile of unmarked papers. After the rating is done by the rater who marks all of the test components, his/her marking is compared to the pre-assessed productions. In the case of each rater, a set of 4 pre-assessed scripts was used and in this way raters were checked for leniency and severity.

Rating of the Speaking component is carried out simultaneously by two assessors who have previously undergone at least one training process. Their only task during the development of the oral examination is to assess the oral production of the candidates using a detailed grid and assessment forms for their own comments regarding the performances. If raters give widely varying marks (a difference of more than 2 points in the final marks), then a discussion takes place during which the examiner can express her opinion on the production of the candidate. If the two raters do not agree, the chief examiner will later grade the oral production as well, as each spoken performance is recorded and can be reassessed anytime.

Rating accuracy was also checked in the case of Writing and Speaking components through qualitative analysis. Rater's comments justifying the points awarded for each criterion in the case of

http://dx.doi.org/10.15405/epsbs.2016.09.132
eISSN: 2357-1330 / Corresponding Author: Lavinia Vasiu
Selection and peer-review under responsibility of the Organizing Committee of the conference

each candidate written on the blank grids (assessment forms) during the session of examination were analysed and compared to the descriptors in the Department assessment grids. If the assessors' comments were not consistent with the descriptors in the Department grids (e.g. the assessor's comments reflect a performance that would be ranked on level 5 according to the descriptors for level 5 in the Department grid, but he/she ranked it on level 4), raters will be closely monitored and, if needed, sent to new training sessions.

Inter- and intra-rater reliability was examined only in the case of the marking of scripts and of spoken performances. For checking **inter-rater reliability** (if different raters rate performances similarly – they do not need to agree completely, but, as they use the same criteria, their ratings should not be wildly different), a correlation coefficient between the two raters was calculated using the Excel Pearson function (in the case of the points given by the two assessors for every criterion for every candidate they both assessed). (see Table 3 below)

**Table 3 .** Points awarded by raters for accuracy in the speaking component. Correlation coefficient

| | Rater 1 | Rater 2 |
|---|---|---|
| Candidate 1 | 5 | 4 |
| Candidate 2 | 4 | 4 |
| Candidate 3 | 4 | 5 |
| Candidate 4 | 3 | 5 |
| Candidate 5 | 2 | 3 |
| Candidate 6 | 1 | 1 |
| Candidate 7 | 3 | 3 |
| Candidate 8 | 3 | 4 |
| Candidate 9 | 4 | 4 |
| Pearson correlation coefficient | 0.73598 | |

If the coefficient is higher than 0.8, the assessors do not need special monitoring (Council of Europe, 2001, p. 79). In case it is lower than 0.8, different measures should be taken, according to the value of the coefficient.

**Intra-rater agreement** (or internal consistency, to what extent do the raters agree with themselves) was checked through both qualitative and quantitative analyses. Raters' comments from the blank grids (assessment forms) on one criterion and for one score (e.g. 5 points) were analyzed and checked for consistency (if the rater uses the same descriptors to judge all the performances he rated with 5 points, for example), then the same was done with the other criteria. If the comments are consistent, the assessor does not need special monitoring. If, however, the comments differ or contradict each other, a range of measures are applicable (feedback sessions, retraining, exclusion from the team, etc.). Also, the performance of raters was monitored by calculating the standard deviation (Excel, STDEV.S function). This, again, was calculated for each criterion (See an example below, in Table 4).

**Table 4 .** Rater 1 – Complexity (speaking component). Average, standard deviation, points awarded

| Average | Standard deviation | No. of candidates with 5 points | No. of candidates with 4 points | No. of candidates with 3 points | No. of candidates with 2 points | No. of candidates with 1 point | No. of candidates with 0 points |
|---|---|---|---|---|---|---|---|
| 4.4 | 0.88257995 | 12 | 5 | 2 | 1 | 0 | 0 |

Then, severity and leniency in the case of each criterion were taken into consideration by checking if there was one criterion where the assessors had the tendency to give more or fewer points (by comparing the number of candidates who got one score for each criterion – Table 5). Those raters who

appeared to be too lenient, too severe or those who show inconsistency in marking or in applying the criteria will be sent to another training session before rating again.

**Table 5** Rater 1 – Speaking component – points awarded (one session)

| Criterion | No. of candidates with 5 points | No. of candidates with 4 points | No. of candidates with 3 points | No. of candidates with 2 points | No. of candidates with 0 points |
|---|---|---|---|---|---|
| Complexity | 10 | 2 | 0 | 0 | 0 |
| Accuracy | 2 | 3 | 4 | 3 | 0 |
| Cohesion and coherence | 10 | 1 | 1 | 0 | 0 |
| Communication efficiency | 12 | 0 | 0 | 0 | 0 |

## 4. Results, discussion and recommendations

As described above, several aspects were taken into account when illustrating the performance of raters: the types and frequency of errors they made and the extent to which they agreed with each other and with themselves. The results demonstrate validity, accuracy and reliability, but they also draw attention on several issues which should be the core of the following training sessions.

### *4.1. Listening, Reading, Elements of Communication Construction (ECC)*

For the first three components of the test, only accuracy of marking could be calculated as they include just items that can be objectively marked.

#### *4.1.1. Error rate*

No error rate was problematic (= higher than 0.02) and there was one case when the error rate was 0 (= one rater who made no mistakes). However, 9 raters did make several mistakes each, so we consider having each paper double-marked a good way of ensuring accuracy in the final score.

#### *4.1.2. Types of errors*

The error rates are not concerning but this does not mean that the errors the markers made cannot reveal some relevant issues in the grading process. Very few errors were made when calculating the final scores for each component or when checking the answers for the true/false, multiple choice or matching items, so, in these cases, it was most probably a focusing problem that caused the very few mistakes. Most of the errors were made when assigning points for the partial credit items (Listening and ECC components). This means that either the marking schemes are not clear enough (= they don't cover all the possible situations) and should be adjusted or that markers were not well enough trained and the following standardization workshops should focus on the process of marking partial credit items.

### *4.2. Speaking and Writing*

In the case of Speaking and Writing, accuracy, inter- and intra-rater agreement, leniency and severity were monitored.

#### *4.2.1. Rating accuracy*

http://dx.doi.org/10.15405/epsbs.2016.09.132
eISSN: 2357-1330 / Corresponding Author: Lavinia Vasiu
Selection and peer-review under responsibility of the Organizing Committee of the conference

Generally, raters used the descriptors from the assessment grids (or references to them) in their comments justifying the points awarded for each criterion. When raters' comments for one criterion mention some aspects that one cannot find in the descriptions, they usually don't seem to have any effect on the score for that criterion, as they are in all cases complemented by other comments in direct relation to the descriptors from the assessment grids. However, there seem to be some features from the descriptors in the Department grids which are preferred by the raters in the case of each criterion. For example, when describing *complexity* in both writing and speaking, most raters use references to complex or simple grammatical structures and vocabulary range is often minimized, although this is also an important aspect of the criterion. When commenting upon *communication efficiency* all raters seem to focus on the extent to which students covered/accomplished the tasks, downplaying other aspects mentioned in the descriptors: how the communicative functions were expressed, problems concerning style and register, etc. Regarding *fluency and coherence* (speaking), most comments refer to the frequency and complexity of the connecting words students use and to the length and frequency of pauses they make. In the case of writing, the criterion is called *Organization* and all raters make comments on layout, connecting words and special formulaic language students use. Very few comments regarding cohesive devices are to be found. When judging *accuracy* all raters seem to stick to the grids in what concerns both writing and speaking. They refer to the types and frequency of errors, to the way errors affect the message and almost all raters write down examples of errors extracted from the students' performances.

Overall, we can conclude that all raters seem to use the assessment grids correctly. Nevertheless, there are aspects which are downplayed in the case of each criterion. Therefore, we consider that this represents an issue that should be discussed during the following standardization workshops which, in our opinion, should be mandatory before each session of examination.

#### *4.2.2. Inter-rater agreement*

The correlation coefficient was calculated for each pair of raters and for each criterion. Only in the case of rating *accuracy* in speaking and *organization* in writing 3 coefficients were lower than 0.8 (for 1 out of 5 pairs in speaking (0.69) and for 2 out of 5 pairs in writing (0.60 and 0.79)). This could indicate the fact that the descriptors for these two criteria (*accuracy* in speaking, *organization* in writing) are not concrete enough or one or both raters from each of the pairs in question need more training. Considering the small number of cases where the coefficient is lower than allowed and the fact that in all these cases the coefficient is not so far from the limit (0.8), we can conclude that raters agree with each other to a high extent.

#### *4.2.3. Intra-rater agreement*

Each rater was monitored for internal consistency by analyzing and comparing his/her comments when assigning a number of points for each criterion. In most of the cases the comments were consistent: the raters used more or less the same words and took into account the same aspects when assigning one score to various performances. There were some isolated cases when the rater had the same comments for a script he/she rated with 4 points and for another one he rated with 5 points, for example, but as these cases were exceptional, they were not considered relevant for the overall performance of the rater.

*4.2.4. Leniency and severity*

In order to monitor the raters for leniency and severity in assigning scores for the writing component, the method of pre-rated scripts was used. Each rater was given 4 pre-rated scripts in the pile of scripts he was supposed to grade. The correlation coefficient (between the number of points assigned by the chief examiner during the pre-rating task and that assigned by the first rater during the rating task) was in all cases higher than 0.78 for the following criteria: complexity and *communication efficiency*. Some problems were identified in relation to the other two criteria: *accuracy* and *organization*. One rater seemed to be too severe regarding grammatical accuracy (he assigned lower scores for all four papers) and two raters proved to be too lenient with respect to the organization criterion (they systematically assigned higher scores than the chief examiner). These results were further confirmed by the other method of monitoring raters for leniency and severity: checking if there is one criterion where the assessors have the tendency to give more or fewer points (comparing the number of candidates who got one score for each criterion). Evidence was found that the same three raters had problems when applying the grid for the two criteria (*accuracy* and *organization*): one of them was too severe (only one script out of 12 received the maximum score for grammatical *accuracy* and 3 out of 12 were rated with 4 points) and two of them were too lenient (they both assigned 5 and 4 points for *organization* for all the 22 papers they graded).

The results of the study demonstrate that the assessors rate accurately, they are consistent with themselves and they agree with other raters to a great extent, they are not too lenient and not too severe. However, the analysis revealed some aspects that could be revised in the grading phase:

- the marking scheme should be adjusted with respect to partial credit items, as some errors occur when rating this kind of items;
- even if raters applied the grid correctly, it seemed that some of the aspects mentioned in the grid were downplayed by the assessors – the following training sessions should focus on raising raters' awareness regarding these aspects (e.g. the *Complexity* criterion refers to both grammatical structures and vocabulary range);
- as some discrepancies between raters were encountered when analysing the points awarded for *accuracy* in speaking and *organization* in writing, the grids could be revisited and enriched with quantitative details (e.g. concrete examples of the connecting words a candidate is expected to use at each band in the grid, type and number of mistakes a student could make, etc.).

## 5. Limitations and further research

As a high-stakes examination, it is important that RFL, level B2 exam continues to demonstrate validity and reliability in all aspects including or especially in what concerns test scores. We believe our small study contributes in some way to the validity argument supporting the use of RFL B2 examinations as means of assessing the communicative competences of those who want to register to any academic program taught in Romanian in any university in our country. Also, we believe that the methods and procedures presented here could raise awareness (in what concerns other examination agencies) regarding the importance of monitoring raters in ensuring accurate, valid and reliable test scores.

*http://dx.doi.org/10.15405/epsbs.2016.09.132*
*eISSN: 2357-1330 / Corresponding Author: Lavinia Vasiu*
*Selection and peer-review under responsibility of the Organizing Committee of the conference*

However, the study should be regarded as a point of departure and the findings are intended to be representative only for one session of examination. Evidently, it should benefit from further analysis – more raters and more sessions of examinations should be observed, comparisons between raters' performances should be made and the evolution/involution of each rater should be analyzed. Also, verbal protocol analysis (Bachman, 1990) could be of great use in the case of rating scripts (raters could be asked to record their thoughts while rating the scripts – the analysis of transcripts could lead to interesting and relevant results).

## References

ALTE Minimum standards for establishing quality profiles in ALTE examinations (2007). Available online at: http://www.alte.org/attachments/files/minimum_standards.pdf

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1985). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Bachman, L. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press. Available online at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Fulcher, G., Davidson, F. (2007). Language Testing and Assessment. An advanced resource book. In Candlin, C., Carter, R. (eds.), Routledge Applied Linguistics, Routledge.

Green, A. (1998). Verbal protocol analysis in language testing research. Cambridge: UCLES/Cambridge University Press.

http://www.alte.org/setting_standards

http://www.alte.org/setting_standards/the_alte_q_mark

Hughes, A. (1989). Testing for Language Teachers, 1st ed. Cambridge: Cambridge University Press.

Manual for Language test development and examining for use with the CEFR – produced by ALTE on behalf of the Language Policy Unit, Council of Europe (2011). Available online at: http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Messik, S. (1989). 'Validity.' In Linn, R. L. (ed.), Educational Measurement. New York: Macmillan/ American Council on Education, 13–103.