STUDIA UBB PHILOLOGIA, LXII, 2, 2017, p. 93 - 112
(RECOMMENDED CITATION)
DOI:10.24193/subbphilo.2017.2.07

STATISTICAL ANALYSIS OF ROMANIAN AS A FOREIGN LANGUAGE PRETEST AND EXAM DATA

LIANA STANCA[1], DINA VÎLCU[2]

ABSTRACT. ***Statistical Analysis of Romanian as a Foreign Language Pretest and Exam Data.*** In the European Union, linguistic policies have gained an essential role within the European policy in its entirety. In this context, in 2014 we started an experiment with the aim of creating an instrument for analysing the quality of tests of Romanian as a foreign language. The instrument thus created contains a succession of statistics tests including factor structure, internal consistency and convergent validity evidence, IRT and ROC analysis for validation. Our creation reached the point where it successfully fulfills the task of Reading the Mind in the process of learning Romanian, also applicable to other languages. The present article shows the manner in which the authors apply the IRT models. It will demonstrate that the IRT models can help improve test scoring and facilitate the development of superior quality tests in Reading Mind applied to the process of studying foreign languages. In the authors' perspective, the present work has a significant contribution to the applicability and usability of IRT models in educational testing, as the presented results show. Furthermore, the paper brings arguments in favour of utilizing IRT and ROC curve models in common, which leads to the conclusion that the achieved results are important in the process of improving the precision of marking and the manner in which the tests are administered through the adaptive use of discriminating elements pertaining to the testing procedure, which saves time, space and impacts on the number of assigned tests.

Keywords: *statistical analysis, language examinations, pretest, IRT model, ROC analysis, quality language tests, reading mind aspects*

[1] Associate Professor, Business Information Systems Department of the Faculty of Economics and Business Administration at the Babeş-Bolyai University, Cluj-Napoca, Romania. liana.stanca@econ.ubbcluj.ro

[2] Lecturer, Ph.D, at the Department of Romanian language, culture and civilisation, Faculty of Letters, Babeş-Bolyai University. She teaches Romanian as a foreign language, working mainly with the students in the preparatory year, but also with other categories of public. She is especially interested in the assessment of Romanian as a foreign language, being involved in the process of creation, administration and marking of the examinations. As a researcher, her interest is mainly directed to the domain of assessment and to that of integral linguistics. dina.vilcu@lett.ubbcluj.ro

REZUMAT. ***Analiza statistică a pretestărilor şi examenelor de română ca limbă străină.*** În Uniunea Europeană politicile lingvistice au câştigat un rol esenţial în cadrul politicii europene în ansamblul său. În acest context, în anul 2014 am demarat un experiment cu scopul de a realiza un instrument de analiză a calităţii testelor de limba română ca limbă străină. Instrumentul care a rezultat este alcătuit dintr-o succesiune de teste statistice, incluzând *factor structure*, *internal consistency*, *convergent validity evidence*, IRT şi ROC *analysis for validation*, acesta ajungând să îndeplinească cu succes rolul de *Reading the Mind* în procesul de învăţare a limbii române şi fiind aplicabil, de asemenea, în cazul altor limbi. Articolul are drept scop prezentarea modului în care autoarele au decis să aplice modelele IRT. Studiul va demonstra faptul că modelele IRT pot ajuta la îmbunătăţirea notării testelor şi pot facilita crearea unor teste de o calitate superioară, care să corespundă tehnicilor Reading Mind aplicate procesului de studiere a limbilor străine. Din punctul de vedere al autoarelor, instrumentul prezentat aici aduce o contribuţie semnificativă la aplicabilitatea şi modul de folosire a modelelor IRT în evaluarea educaţională, aşa cum va reieşi din rezultatele prezentate. În plus, în lucrare se aduc argumente pentru utilizarea în comun a modelelor IRT şi ROC *curve*, concluzionându-se că rezultatele la care s-a ajuns sunt importante în procesul de îmbunătăţire atât a preciziei în notare, cât şi a modului în care se administrează testele, prin utilizarea adaptivă a elementelor discriminative ale procedurii de testare, acest proces generând o economie de timp, spaţiu şi număr de teste administrate.

Cuvinte cheie: *analiză statistică, examene de competenţă lingvistică, pretestare, modelul IRT, analiza ROC, teste de calitate, aspecte reading mind*

1. Statistical analysis. What it stands (and does not stand) for

Among other major things, the world we live in is one in which language abilities have become vital for our social and economic life. The role of language testing has never ceased to grow, making, in exchange, its way towards the central part of our professional, educational and social life. In this context, the need of trustworthiness has only been placing higher and higher pressure on the language tests and on their providers. However, the domain of language testing was never far behind the trends of progress in sciences and technology, keeping pace with new methods in the field of social sciences, adapting and using instruments of measurement, which allowed language testers to ensure objectivity and fairness, validity and impact of their products. Psychometric analysis started to be used in language testing in the 60s, when the most modern model of language assessment was known as 'structuralist-psychometric'. Associated with the format of test based on

discrete items, this model was given credit for objectivity, reliability and validity (Hawkey: 21). Even if things have changed tremendously in time, McNamara shows that practices adopted at that time remained highly influential (McNamara: 14). Statistical analysis, a significant part of psychometrics, which will be - related to tests of Romanian as a foreign language - the main concern of this study, has seen, in the same period, "the blossoming and refinement of techniques for examining many variables simultaneously" (Bachman 1990: 296). Since those times, the importance and influence of statistical analysis in language testing has constantly grown.

Statistics, "a set of logical and mathematical procedures for analyzing quantitative data" (Bachman 2004: 3), has become a tool without which almost no institution of language testing can imagine working nowadays. An example comes from the Cambridge examinations, in relation with which Geranpayeh says: "Measurement theory and practice are now so embedded in what Cambridge English does that it is hard to think of a single process within which is not impacted by them" (Geranpayeh: 4). Some of these processes are as follows: 1) statistical analysis of objective items, which shows how well an item works and also indicates the problems which need to be addressed. The facility value and the discrimination index are two of the most important aspects to be calculated for each objectively marked item. *Item analysis*, together with its interpretation, is, according to Corrigan and Crump, "a cornerstone in appraising item quality and making decisions concerning the use of items in future live tests" (Corrigan, Crump: 4). The facility value and the discrimination index are population dependent (Relating language examinations: 94). Being related to the bio data of the test takers, a correct calculation of these values will ensure fairness to the test taking population and will vouch the proposed items for the live examinations. 2) While construct comparability between different versions of the tests is based mainly on test specifications, *psychometric comparability* is another aspect which can benefit from statistical analysis (Docherty, Corkill: 11). 3) The *marking and grading* of the tests, as well as the *monitoring of assessors* are also subjected to statistical analysis, which provides proof of objectivity, fairness and consistency. 4) *Interpretation of test results*, based on *grading decisions*, impacts directly on the use of the tests "allowing certificate end users, for example potential employers or universities, to draw appropriate inferences and make decisions where appropriate. This is a central aspect of the testing process and one on which its usefulness hinges" (Elliott, Stevenson: 14). Thus, statistical analysis 'has infiltrated' the domain of language testing to the point where designing items and tests, marking and deciding upon results without the use of this tool has become unconceivable. On condition of a fair and correct use of statistical analysis, this can only have a positive impact on objectivity, fairness, validity and reliability of

examinations. The examinations of Romanian as a foreign language applied at the Department of Romanian language, culture and civilization, Babeş-Bolyai University, benefit now from statistical analysis at all the levels presented above, an aspect which, corroborated with the qualitative aspects of the whole process of assessment, increased their validity and reliability to the point of excellence.

However, language test specialists have often pointed out the danger of overemphasizing the importance of statistical analysis over aspects that are also crucial in all the processes of language testing. Thus, according to Cyril Weir, statistical data do not in themselves generate conceptual labels. We can never escape from *the need to define what is being measured* (italics mine), just as we are obliged to investigate how adequate a test is in operation (Weir: 14). Alderson also draws attention to what assessing language involves, and this is not only the technical skills and knowledge to construct and analyse a test – the psychometric and statistical side to language testing – but also a knowledge of *what language is*, *what it means to 'know' a language*, and *what is involved in learning a language* (italics mine) as your mother tongue or as a second or subsequent language, what is required to get somebody to perform, using language, to the best of their ability (Alderson: 12). In the end, what we need is, according to Bachman, a balance "between language learning (theories) and measurement in language testing: We language testers thus cannot allow ourselves the delusion that current views of language use and language learning can be ignored because this simplifies the problem of measurement. Nor can we afford the luxury of placing all our confidence in the correctness of our applied linguistic theories, in the hope that the measurement issues will thus evaporate. Progress does not lie in the direction of more sophisticated psychometric analyses of theoretically unjustified tests. Nor will we advance if we ignore fundamental measurement considerations in our striving to capture authenticity in our testing procedures. The challenge will be to address the perennial dilemmas of language testing with both an open mind and a healthy respect for the complexity of the abilities we want to measure, and a continued commitment to utilize the best and most appropriate psychometric theories and statistical tools at our disposal so as to assure that we can be held accountable for the uses we and others make of our tests" (Bachman 1990: 352-3).

2. The case of Romanian as a foreign language (RFL)

The process of assessing RFL started to change, at the Department of Romanian language, culture and civilization from Babeş-Bolyai University in Cluj-Napoca, around 2005. We could also say that it culminated with the years 2013-2015, when the department intensified the preparations for the ALTE

(Association of Language Testers in Europe) audit. Our department submitted a suit of examinations (A1, A2, B1 and B2) to the audit with the purpose of having the quality of our examinations confirmed and becoming, based on the outcome of this audit, members of ALTE.

The clear and stable structure of the evaluation system at our department, which received its present form after consistent and continuous improvement in the period indicated above, was the incentive for the creation and publication of testing materials, for the constant enhancement of the process of item creation and of that of test administration, in the marking and grading of results and in monitoring of assessors (Vîlcu 2014: passim). All along the way, the teaching system was progressively modelled after the evaluation one and after materials designed at the Department.

The second part of this study will be dedicated to statistical item analysis, with a description of the instruments and procedures we apply in pretesting and in examinations. Consequently, we will exemplify here with only one situation the way in which psychometric analysis is involved in other stages of testing, in our examinations. We will also try to illustrate the challenges that this type of exam data processing placed on assessment of Romanian as a foreign language at our department.

2.1. The constitution of the item bank. The basket method

The items which get to be actually answered by the student in live examination cannot just get in the exam paper directly from the item writer who produced them. They need to run through a process of validation before being used (Vîlcu 2015: 71-72) and need also to be processed after they have been used in live examinations. In our case, this process was developed as follows:

- based on specifications we had previously designed, a team of item writers prepared a large number of items for all the components of our examinations (listening, reading, elements of communication construction, writing, speaking);
- the items were revised within our work team, with feedback from the members of the team;
- we submitted the items to the analysis of a panel of experts, with two purposes: 1) to have the confirmation of the level for each item of every task we proposed; 2) to collect observations from the experts in order to adjust the quality of the items and of the task input (I am going to present here only the part of level appreciation by the experts, the description of the processing of the observations related to the quality of items is shortly described in Vîlcu 2015: 71-72);

- for having the confirmation of the level of the items, we chose to use the basket method (Manual for relating: 75-77); we selected the members of the panel of experts and we invited them to participate in this project;
- we invited the selected experts to participate in a standardisation workshop we organised at the Department, in order for them to re-familiarise with the CEFR documents and to be presented the new assessment instruments we had devised at the Department;
- we placed the items we had produced in separate documents, for each component, in order, from the simplest to the most complicated (from A1 to B2), without giving any indication on the level we had created the items for or on the number of tasks and items we included for every level; we created a document for the experts in which we kindly asked them to place each of the items of each of the tasks on one level (between A1 and B2; if they considered that there were items more difficult than B2, they could write *more difficult than B2*, without having to further specify a level for these cases); the tasks were numbered in order, from the first to the last in the document, without considering the level they had been created for; we placed an empty box next to each of the items of each of the tasks, where the experts were going to write the level they considered that item was adequate for; we placed one empty box for each task, where the experts could comment on other aspects related to the quality of the task/of all items or some items of the task (e.g. the quality of the text used as input);
- we registered the experts' answers concerning the level of the items in an excel document;
- we processed the appreciation of the level by the experts by calculating the degree of deviation for each item; the deviation was calculated first uni-directionally (identifying the items whose level appreciated by the experts was higher than the one we created the item for); the calculations were performed as follows: if, for example, the initial level, the one proposed by the item creators, was A1 and the expert decided on A2, there was one point of deviation, if the expert decided on B1, then there were two points of deviation, etc.; then, the total of the deviation points was divided to the number of the experts who estimated the level (in the case of some of the items, there were experts who did not express any estimation); if the expert gave an estimation which covered two levels (e.g. A1/A2), then one point and a half of deviation was calculated;
- the deviation was then calculated bi-directionally (identifying the items whose level appreciated by the experts was both below and higher than the one we had created the item for); the evaluations below the initial level and the ones higher than the initial level for the same item were not

compensated; the points of deviation in both directions were summed up and then the total number of point was divided to the number of experts who expressed their opinion in relation to that item (Table 1);

- the revision of the items was performed in accordance with the results from the bi-directional calculation of the deviation; a maximum value of deviation of 0.50 points was accepted; after the revision, the items were included into an item bank, from where they are selected for pretesting; only after confirmation of their quality in pretesting or after necessary revision do these items find their way and their place in live examinations.

Table 1. Calculation of deviation for proposed items

	A	B	C	D	E	F	G	H
1		task 1.1	task 1.4	task 1.5	task 2.2	task 2.3	task 2.5	task 3.1
2	Expert 1	A1	A1	A1	A1	A1	A1	A1
3	Expert 2	A1	A1	A1	A1	A1	A1	A1
4	Expert 3	A1	A1	A1	A1	A1	A1	A1
5	Expert 4	A1	A1	A1	A1	A1	A2	A1
6	Expert 5	A1	A1	A2	A2	X	A1	A2/B1
7	Expert 6	A1	A1	A1	A1	A1	A1	A1
8	Expert 7	A1	A1	A1	A1	A1	A1	A2
9	Expert 8	A1	A2	A2	A1	A1	A1	A2
10	Expert 9	A1	A1	A1	A1	A1	A1	A1
11	Expert 10	A1	A1	A1	A1	A1	A1	A1
12			0,1	0,2	0,1	0	0,1	0,35

2.2. Challenges for RFL examinations

2.2.1. Romanian as a foreign language has been taught and assessed in Cluj-Napoca and in the rest of the country for decades, to thousands of students yearly. However, a preoccupation for the necessary modernisation and promoting of this domain has been noticed rather in the last years. All along the way of the process of preparation for the ALTE audit, this situation placed us in the position in which we had to adopt and adapt new procedures (e.g. item banking, statistical analysis of the items in relation with bio data, monitoring raters (Vasiu, Arieșan: passim), etc.), to strengthen the relating of assessment to the CEFR or to design new materials, theoretical and practical, to explain the fundaments of our testing system and also to provide this system with the necessary instruments for

assessment (*Descrierea minimală a limbii române*, *Manual de limba română ca limbă străină. A1-A2*, numerous and various studies dedicated to the process of teaching and assessing RFL, instruments of assessment - grids, samples of assessed oral and written productions), etc. The main points of reference in the new design of the RFL examinations were the systems created for renowned examinations for other languages and the literature in the field. One of the most problematic instruments to develop for our examinations was the system of statistical item analysis. The specialists in our university needed to familiarise with this new field of analysis and to adapt and/or create sub-systems of calculation perfectly adequate to item analysis in RFL exams. However, the statistical analysis has developed into an ongoing process which applies to our examinations and contributes to the relevance of the results we issue.

2.2.2. A problem which we needed to overcome in the development of a new, modern and relevant assessment system for RFL was the fact that in present, unlike decades ago, we have, at Babeş-Bolyai University, a rather small number of foreign students. We developed the system of statistical analysis based on the results of the students in the programme of preparatory year (language year) due to the fact that they have the longest programme of learning Romanian, covering the levels A1 to B2. In present, we have approximately 120 students per year, who get through all these language examinations. Consulting the specialised literature, we concluded that this number would be considered rather small for relevant results of statistical analysis (Manual for relating: 94). In order to compensate for this shortcoming, we decided on a number of measures to make our results as relevant as possible. Thus, we decided to apply statistical analysis both to the items in the examinations, and the ones used in pretesting. The results of the statistical analysis after pretesting give us the chance to improve our items before introducing them into live examinations. Another solution we found was to combine the quantitative methods with the qualitative ones. We apply questionnaires at the end of every examination and we correlate their results with the ones we obtain from the statistical analysis. A third method that we use is the application of statistical analysis every time an item is used both in pretesting and in examinations. The results that are obtained by the test takers are compared and this way the items get through successive stages of validation. For this procedure, the fact that our group of subjects is rather homogeneous from one year to the next is really helpful.

The efforts we have made to overcome the difficulties we had in creating a system of assessment of RFL in our university were significant. However, they only contributed to making this system valid and highly reliable.

3. Materials and Method

3.1. Description of the method

The method employed for collecting data in our research is the test. At the end of a 90-120 hour course the students receive and sit an evaluation testing their acquired knowledge. The target audience is represented by students who attend courses where they are taught Romanian. The instrument for measuring their knowledge of the Romanian language is composed of items that have either the 5 step Likert scale or the dichotomous method as an answer. We used the MS Excel application for collecting data in our statistical analysis, whereas for the statistical processing we used the following softs: SPSS13 (Statistical Package for the Social Sciences) for Windows, Statistica 7.

The general hypothesis of the study is the Romanian language test, which is a valid instrument, having the power to accurately discriminate the level of knowledge of the person taking the test.

In the process of validating the battery of tests that has the purpose to identify the level of knowledge of the Romanian language acquired by the candidates, we had to create a statistical procedure that would contain the descriptive analysis of data, Item Analysis, Roc-Curve, and cluster analysis. Our demarche was achieved on basis of the already existent results in the specialized literature provided in this field (Baker, 1992), (Birnbaum, 1968) (Bock, 1972). These authors consider that the analysis of the items represents a collection of statistical procedures which allow the description of the relationship between the items of a test as well as that of the relationship between each item and the overall score. The result of the enterprise is a statistical procedure built with the purpose of validating the test as an assessment and testing instrument of the level of knowledge in the field of the Romanian language as well as the aim of creating clusters of participants at the Romanian language course. The steps of the statistical procedure represent an extrapolation of the created nucleus, as it is defined by (Ellis & Mead, 2004), containing the following steps:

1. The descriptive analysis of the items which emphasize the calculation of the difficulty index or parameter of the items. The difficulty index (*Imran Zafar. 2008*) was used to measure how easy or how difficult each item in our test actually is. The reduced value tending towards 0 of the index of difficulty indicates that the item is difficult (Aiken1994).
2. The evaluation of the capacity to discriminate between the items. The discriminant parameter has the role to measure the level of knowledge of the persons sitting the test. The discriminant index (*Imran Zafar. 2008*) is strictly connected to the difficulty parameter characteristic of the items. A

low difficulty item is solved by the majority of the students, so the discriminating index tends towards 0. This means that it is poor, but if an item with an increased level of difficulty implies an increased level of knowledge for the student and implicitly the discriminating index tends towards 1, then this means that it is *rich (Imran Zafar. 2008)*.

3. The evaluation of the internal consistency of the battery formed by tests which are destined for the evaluation of the level of knowledge of the Romanian language was analysed. This supposes the creation of a series of successive evaluations of both the relationships among items as well as the relationship between items and the global score. The purpose of these evaluations is the selection of the relevant items according to their relationship with the global score. The main criterion for this operation is the value of the Cronbach alfa parameter, which has a variation field of 0 to 1. According to the literature in the field (Crocker, Algina,1986), a scale can be considered consistent in case the value of the Cronbach alfa index is as close to 1 as possible, the level of 0.70 being accepted, as a convention as the minimum limit (McDonald, R. P. 1999). In our study, we have calculated the Cronbach's Alpha Coefficient with the purpose of determining the reliability of the test batteries destined to the correct identification of the level of knowledge of the Romanian language that each participant had. At this stage we had to apply the test measuring validity and reliability. Through reliability we measured the accuracy, stability and coherence of the test results. The reliability (McDonald, R. P. 1999) is a test which is influenced by the characteristics of the students, by those of the test and also by the conditions in which it is administered, tested and graded. The reliability of a test refers to the measure in which the test is able to produce consistent results (McDonald, R. P. 1999), (Crocker, Algina, 1986). The validity of the test allows us to establish and present the level of adequacy of the given test for the studied area. The internal consistency of the test, measured through split-half method, was found to be good (coefficient Guttman split-half, Spearman-Brown). According to the Spearman-Brown results, the number of test items is adequate, no increase in their number is necessary; however, if this happened, the test reliability would not be significantly modified. (McDonald, R. P. 1999) (Crocker, Algina, 1986). The internal consistency of the test was also calculated and the Rasch analysis was used in order to confirm that the test measures the person's capacity to use Romanian. The **Rasch model** of validation of statistic tests **(**Bond TG, Fox CM.2001**),** **(**Slinde, J, Linn, R.**)** was put in practice. The null hypothesis was that all the elements are equally discriminating. The *Two-Parameter Logistic model*

(2PL) was applied, considering two parameters: difficulty and student ability. The result was that the test is composed of simple, moderate and difficult items. In the test there are items which ensure discrimination between the students who answer the questions by chance or knowingly. The result was that the test assesses the language abilities of the persons as a whole.

4. The evaluation of the single dimension of the test through which we checked whether the items which form the test in the Romanian language only cover one dimension, a single latent factor. This basic postulate is known under the term of one-dimensionality or single-dimensionality and is linked to a second assumption, namely that of the local indepenence of items. According to (Hambleton, Swaminathan, & Rogers, 1991), it is mentioned that one-dimensionality does not strictly refer to the presence of a single dimension, but to the existence of a dominant dimension that influences the performance in a test. This dominant dimension is called ability or, more generally, coverage in the latent factor.
5. We achieved the evaluation of the performance model through the receiver operating characteristic curve (Mason & Graham 2002). The ROC analysis was used with the purpose of checking the power of discrimination that the test had. The result of this test offers us the possibility to conclude on whether the organization of the Romanian language test is a discriminative model or not.

3.2. Processed data and results

In what follows we shall present the results that we obtained in *2013-2014. The total number of participants is of 195 in the A1, A2, B1 and B2 levels of language testing.*

The descriptive analysis of the data considered in the study offered us an image on the profile of the person who studies Romanian: reason/ purpose, time allotted to studying, sex, profession, nationality. For example, in the study of 2013 applied to the A2 level, the obtained results are as follows:

1. An average number of months allotted to learning Romanian is high for the persons who have the purpose of learning Romanian in order to study here, when compared to the persons who learn Romanian for personal reasons, and ultimately when compared to those people who learn for professional purposes.
2. The persons considered in the study have the average age of 21.96 with a standard deviation of 5.876, 95% CI (20.21; 23.70). The minimum age is 18 and the maximum is 44.

3. The persons who take part in the study allow an average of 20.65 hours of learning Romanian, with a standard deviation of 3.267,95% CI (19.68;21.62).
4. The participants know at least one other foreign language (the greatest majority speaks English), apart from their mother tongue and Romanian, having an average level of foreign language knowledge of A2.

The summary of the results obtained in the descriptive analysis stage for each level is as follows:

Level A1: number of participants 54; medium age: 21,80; more than 50% between 18 and 20 years old; almost 70% men; **occupation:** more than 96% students; **studies:** almost 70% high-school graduates; **mother tongue:** more speakers of Arabic and Albanian; **how long they studied Romanian:** more than 50% - 2 months (160 hours); **why they study Romanian:** more than 85% - for studies; **other known languages**: more than 50% speak another language at intermediate level; out of these, more than 50% speak English.

The scenario of the second part of our statistical study is the following: descriptive analysis with coefficient of difficulty for each item of the test, verification of survey's validation, followed by reliability test. Item difficulty is calculated. The minimum value is 1, the maximum value is 3 for Reading and Listening with the minimum value being 1 and the maximum value being 5, for structural competence (elements of communication construction). The elements with p value over 0.90 and under 0.20 were or are to be carefully revised. Some of them are totally replaced and others are modified. The linguistic abilities of the candidates must be moderate to low, usually speaking. The conclusion, for all the levels, for listening, reading and elements of communication construction is that the difficulty of the items is adequate for the knowledge and competence they are meant to measure. Factor analysis is used in both cases, on data reduction and research and it validates the studied issue, namely Internal Consistency Analysis. The analysis consists of Test Chi-Square, the Anova Test, based on Shapiro-Wilk normality test in order to realize the comparisons imposed by the study (Drugan Tigan 2005). We got the following results:

1. The first step of the statistical analysis consisted of the survey's accuracy verification based on the collected data, approach that imposed the following statistical analysis:

1. A1. Level Listening Test

1.1. The Internal Consistency analysis method allowed us to calculate the internal consistency coefficient Cronbach's alpha = 0.882, which indicates a unitary moderate structure of the used tool, but one that is sensitive on the measured characteristics, which can provide a correct overview of the statistical analysis. The internal consistency of the test, measured using split-half method is

good (Guttman split-half coefficient = 0, 759, Spearman-Brown r = 0,863). Thus, we can observe that the loyalty test slightly changes, according to the Spearman-Brown result for loyalty coefficient. The number of test items is appropriate, it does not impose an increase, but if this happens, the test's loyalty would not change noticeably. At this point we calculated the discrimination coefficients. We calculate the coefficient of discrimination for the items in the same three components of the pre-testing. Therefore, the items with the result poor discrimination are revised.

1.2. The outcome confirms that the survey's questions tended to belong to the studied area. If the current study was re-applied, students' appreciations would indicate a little change compared to this test. The study's hypothesis is unilateral, according to the result of the analysis of different groups and interaction effect with the tool made with Anova (Chi-Square=26.680, p-value=0.000). The Anova test as a tool for validating knowledge or validating individuals' overall test can more accurately differentiate between people who merely guess the answer and those who were ready and had thoroughly learnt for the test. The null hypothesis (There is no limited set of factors which determines the students' level of knowledge) is rejected after the ANOVA test and factor analysis and the alternative hypothesis is accepted. According to this, there is a limited set of factors which determines the validity of the instrument.

1.3. The analysis continued with factor analysis. The first step in such a study is to eliminate an item which shows the correlation smaller than 0.3. Within the study there were no such items. Factor analysis is appropriate for our model, confirmed by the result of the Kaiser-Meyer-Olkin test (=0.717), which specifies how data variability is caused by the tool.

The results of the stage destined for the identification of item discrimination capacity are the following:

1. the test contains items with a high power of discrimination and also items with a moderate to inexistent power of discrimination;
2. the study concludes that the test can be considered an instrument for verifying the students' knowledge, for a group of students with the same level of knowledge;

The results obtained in the stage of calculating the item difficulty coefficient are as follows:

1. from the analysis of the values of the items difficulty index, we conclude that the majority of the items in the study have the degree of difficulty moderate to easy;
2. the linguistc abilities of the candidates must be moderate to high.

2. The A1 level test for Reading Assessment

2.1. The Internal consistency analysis method allowed us to calculate the internal consistency coefficient according to Cronbach's alpha = 0.802,

which indicates a unitary moderate structure of the used tool, but one that is sensitive on the measured characteristics which can provide a correct overview of the statistical analysis. The internal consistency of the test measured using split-half method is good (Guttman split-half coefficient = 0,878, Spearman-Brown r= 0,935). For the dichotomous data, the level of validity and fidelity was calculated using Kuder-Richardson (KR20) (*Imran Zafar. 2008)*. The result in our study is r_{KR20}.=0.972, which means that the level of fidelity of the test items is high. More exactly, it can find and trace the knowledge of the tested persons in that particular domain. The result of the internal consistency coefficient Kuder-Richardson (KR21) is r_{KR21}.=0.964, which implies that the internal consitency of the items measuring the knowledge of that person is very good. On the basis of the results obtained in this stage, we came to the conclusion that the 20 items which form the test and which are destined to testing the level of knowledge of the Romanian language can be used in the testing of persons who are not speakers of our language. As a result, the test has a high level of reliability, demonstrating that the questions in the test had the tendency to form a unity.

2.2. The outcome confirms that the survey's questions tended to belong to the studied area. If the current study were to be re-applied, students' appreciations would indicate a little change compared to this test. Our study's hypothesis is unilateral according to the result of the analysis of different groups and interaction effect with the tool made with Anova (F=4.523, p-value=0.000). Within this process the Levente p=0.000 test rejects the idea that students answer homogenously. The study continued with the discriminative analysis as a predictor: the test can differentiate between the students who learnt and those who didn't. The Wilks parameter is statistically significative (χ^2=76.202,p=0.000). The analysis goes to show that there is a single discriminative function according to which one can differentiate between the students who learnt and those who did not. In conclusion:

1. the number of items is adequate; if the number of items were increased, the reliability of the test would not be significantly modified.
2. reliability is good: the questions in the test tend to constitute a whole.

The results in the stage destined to identify item discrimination capacity are as follows:

1. if a parallel test were developed with similar elements, the relative scores obtained by the students would reflect a small change in comparison with the present test.
2. the test contains items with a high power of discrimination and also items with a moderate to inexistent power of discrimination;

3. the study concludes that the test can be considered an instrument for verifying the students' knowledge, for a group of students with the same level of knowledge;
4. the items whose discrimination type is *poor item* need to be revised (I1, I5).

The results of the stage destined to calculating the item difficulty coefficient, constituted in accordance with (Ghiselli, Campbell, and Zedek, 1981), are the following:

1. there are 6 items with a low level of difficulty, while the others demand higher abilities for being answered;
2. the linguistic abilities of the candidates must be moderate to low;
3. the elements with p values of over 0.90 and under 0.20 necessitate careful evaluation (I1, I4, I5, I8-I11, I16, I17).

4. A1 Level Test Applied to Elements of Communication Construction

4.1. The Internal consistency analysis method allowed us to calculate the internal consistency coefficient Cronbach's alpha = 0.902, which indicates a unitary good structure of the used tool, but one that is sensitive on the measured characteristics. This can provide a correct overview of the statistical analysis. The internal consistency of the test, measured using split-half method is good (Guttman split-half coefficient = 0,818, Spearman-Brown r= 0,901). The conclusion is that the test exhibits a high level of reliability, which means that the questions in the test have the tendency to form a whole. The students who answered correctly at a given question have more chances to answer other similar questions in the same correct way. If a parallel test were created with similar elements, the relative scores of the students would show a slight change when compared to the given present test.

4.2. The outcome confirms that the survey's questions tended to belong to the studied area. If the current study was re-applied, students' appreciations would indicate a little change compared to this test. Our study's hypothesis is unilateral, according to the result of the analysis of different groups and interaction effect with the tool made with Anova (Chi-Square=11.842, p-value=0.000). The Anova test as a tool for validating knowledge or validating individuals' overall test can more accurately differentiate between people who guess the answer and those who were thoroughly prepared and had learned for test in a serious manner. The null hypothesis (There is no limited set of factors which determines the students' level of knowledge.) is rejected after the ANOVA test and factor analysis, so that the alternative hypothesis is accepted. According to this, there is a limited set of factors which determines the validity of the instrument.

4.3. The analysis continued with factor analysis. The first step in such a study is to eliminate an item which shows the correlation smaller than 0.3. Within the study there were no such items. Factor analysis is appropriate for our model, confirmed by the result of the Kaiser-Meyer-Olkin test (=0.303), which specifies how data variability is caused by the tool. The validity of instrument in case of Elements of Communication Construction is as follows:

- the test contains simple, moderate and difficult items; the correct response to items necessitates a medium to high level of knowledge;
- from the analysis of the values of the items difficulty index, we conclude that the majority of the items in the study have the degree of difficulty moderate/optimum; there is a small percentage of difficult items and a small percentage of easy items;

The results obtained in the stage destined to identifying item discrimination capacity are:

1. the test contains items with a high power of discrimination and also items with a moderate to inexistent power of discrimination;
2. the study concludes that the test can be considered an instrument for verifying the students' knowledge, for a group of students with the same level of knowledge;

The results obtained in the stage aimed at calculating the item difficulty coefficient, constituted in accordance with the findings of (Ghiselli, Campbell, and Zedek, 1981), are the following:

1. the test contains simple, moderate and difficult items; the correct response to items necessitates a medium to high level of knowledge;
2. from the analysis of the values of the item difficulty index, we conclude that the majority of the items in the study have the degree of difficulty moderate/optimum, and there is a small percentage of difficult items and a small percentage of easy items as well.

The following step in the statistical analysis is aimed at applying the Roc-Curve test with the purpose of checking the power of discrimination of the test. The result of this Roc-Curve test offers us the possibility to conclude whether the organization of the Romanian language test is a discriminative model or not. In the case of the A1 level test we have provided the following interpretation according to the results offered by statistical analysis: the test presents a high level of reliability, which demonstrates that the questions within the test have the tendency to form a unitary whole. The students who answered one question correctly had more chances to answer other questions in the same way. If a parallel test were created with the help of similar elements, the students' relative scores would exhibit a slight change when compared to the present test. The obtained results in the case of the A1 level test are the following:

- For listening: the studied model is a discriminative one that is perfect for 100% of the cases, so random guessing would register an AUC of approximately 0,247.
- For reading: the studied model is a discriminative one perfect for 100% of the cases, so random guessing would register an AUC of approximately 0,877.
- For elements of communication construction: the studied model is a discriminative one perfect for 100% of the cases, so random guessing would register an AUC of approximately 3.241.

The fourth stage is dedicated to cluster analysis. **Cluster analysis (hierarhical and k-means method)** is applied in order to determine if there are differences between groups of students with different scores. The grades of the students were used as a variable. The conclusion of the study is that the indicators which were included in the study can be considered important for determining if a student has or does not have a certain level in using Romanian appropriately.

The last step is dedicated to the Rasch model, (Bond TG, Fox CM**).** The Rasch model of validation of statistic tests was used in order to verify the results obtained through the above-mentioned statistical procedure. The null hypothesis was that all the elements are equally discriminative. The *Two-Parameter Logistic model* (2PL), considering two parameters: difficulty and student ability, was applied. The result was that the test is composed of simple, moderate and difficult items. In the test there are items which ensure discrimination between the students who answer the questions by chance or knowingly. The result was that the test assesses the language abilities of the persons.

According to the results, the tests contain items with a high power of discrimination and also items with a moderate to inexistent power of discrimination. The analyses conclude that the tests can be considered to be appropriate instruments for assessing the candidates' knowledge and competences.

The above-mentioned test was applied also to testing the A2, B1 and B2 levels of language. The obtained results suggest that the tests have a high level of reliability, a validity that proves that the questions in the test had the tendency to form a whole. The students who answered one question correctly had the chance to answer more questions correctly. If a parallel test were created with the help of similar elements, the relative scores of the students would present a slight change when compared to the present test.

Conclusion

Developing a test that is perfect to assess the level of knowledge in the field of foreign language study is the impossible objective of anyone who is in the position of the evaluator. The statistical procedure we created gravitated

around the difficulty index and the discriminative nature of items within the test. The purpose of the statistical procedure presented in the given material is the significant improvement of the quality of exams that identify the level of knowledge of the Romanian language.

This article creates a general representation of the necessary stages to develop batteries of tests destined to test the levels of mastering foreign languages. The correctly written tests are those that have the role to basically read the mind. This is to say that we need to identify and determine whether the representation of a complex system of words and rules in one single brain of a student is sufficiently similar with a representation of the same system in other brains of other students. Such tests can be viewed as an instrument of checking the knowledge of students belonging to a group that has the same level of knowledge, the same way of thinking.

More limitations would restrict the generalization of our results. First of all, the psychometrical instrument created in this study has carefully selected items, so that in the future we intend to render it better in the domain of functional aspects meant at identifying aptitudes and abilities in order to inform and create a diagnose or in order to predict the performance of individuals applied to real tasks assigned to the participants of this test. The future studies in this domain might eliminate this shortcoming. In the following years, re-applying the batteries and applying statistical analysis as it is presented in this article would aim at refining the instrument in order to obtain a higher level of accuracy in the process of evaluating the level of knowledge and mastering of the Romanian language, irrespective of the mind mapping of each participant.

BIBLIOGRAPHY

Charles Alderson, "The shape of things to come: will it be the normal distribution?" in *European language testing in a global context. Proceedings of the ALTE Barcelona Conference*, 2001, p. 1-27.

Lyle F. Bachman, *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press, 1990.

Lyle F. Bachman, *Statistical Analyses for Language Assessment*, Cambridge, Cambridge University Press, 2004.

Michael Corrigan, Paul Crump, "Item Analysis", in *Research Notes*, issue 59, 2015, p. 4-9.

Coreen Docherty, David Corkill, "Test construction: The Cambridge English Approach", in *Research Notes*, issue 59, 2015, p. 10-14.

Mark Elliott, Lynne Stevenson, "Grading and test equating", in *Research Notes*, issue 59, 2015, p. 14-20.

Ardeshir Geranpayeh, "Introduction", in *Research Notes*, issue 59, 2015, p. 3-4.

Manual for *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* accompanied by *Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*, Strasbourg, Council of Europe, Language Policy Division, 2009.

http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Tim McNamara, "Language Testing", Oxford: Oxford University Press, 2000.

Roger Hawkey, "A Modular Approach to Testing English Language Skills. The development of the Certificates in English Language Skills (CELS) examinations", Cambridge, Cambridge University Press, 2005.

Vasiu Lavinia-Iunia, Arieşan Antonela, 2016, "The Role of Monitoring Raters in Ensuring Accurate and Meaningful Test Scores. Case Study: RFL Examinations", în *WLC 2016 - World LUMEN Congress. Logos Universality Mentality Education Novelty 2016 LUMEN 15th Anniversary Edition*, Volume XV, p. 1068-1076, (coord. Antonio SANDU, Tomita CIULEI & Ana FRUNZA).

Dina Vîlcu, "Relaţionarea examenelor de limba română ca limbă străină la CECR", în Elena Platon, Antonela Arieşan (ed.), *40 de ani de limba română ca limbă străină la UBB. 1974-2014*, Cluj-Napoca, Ed. Casa Cărţii de Ştiinţă, 2014, p. 20-29.

Dina Vîlcu, "Romanian as a foreign language in the context of language assessment in Europe", în *Studia Universitatis Babeş-Bolyai*, *seria Philologia*, nr 2/2015, p. 65-76.

Cyril Weir, "Three lessons from the historiography of language testing", 2014 CRELLA Summer Research Seminar, https://www.beds.ac.uk/crella/seminars/crella-research-seminars/crella-summer-research-seminar-2014

Aiken, Lewis R. 1994. *Psychological Testing and Assessment*, (Eight Edition), Boston: Allyn and Bacon.

Bond TG, Fox CM. *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum Associates; 2001.

MAT S. Deviant, *The Practically cheating statistics handbook*, ISBN-13: 978-1449957858

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Dekker.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability". În F. Lord, & M. Novick, Statistical theories of mental test scores. Reading, MA: Addison Wesley.

Bock, R. (1972). "Estimating item parameters and latent ablility when responses are scored in two or more nominal categories". *Psychometrika* (37), 29-51.

Crocker, L., & Algina, J. (1986)."Introduction to classical & modern test theory". Orlando, FL: Holt, Rinehart and Winston

Drugan T., Achimas A., *Tigan S. Biostatistică*, Editura SRIMA,Cluj-Napoca , ISBN:973-85285-5-0. (2005)

Hambleton, R., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of Item Response Theory*. London: Sage Publications Inc.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman. Google Scholar

Imran Zafar. 2008. "Item Analysis Assumptions, Measure of exam internal consistency (reliability) Kuder-Richardson 20 (KR20)". Artikel Juni 2008. Database Administrator, Assessment Unit, Dept of Medical Education. Ext. 47142

Mason, S. J. and Graham, N. E. (2002), "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation". Q.J.R. Meteorol. Soc., 128: 2145–2166. doi:10.1256/003590002320603584

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Slinde, J, Linn, R. "The Rasch Model, Objective Measurement,Equating, and Robustness". visit at 19.02.2016 -conservancy.umn.edu/bitstream/handle/11299/99830/1/v03n4p437.pdf

*** Item Response Theory: Simple Definition was last modified, March 14th, 2017, www.statisticshowto.com/item-response-theory/

*** Rasch Model/Rasch Analysis: Definition, Examples, www.statisticshowto.com/rasch-model/