STUDIA UBB PHILOLOGIA, LXII, 2, 2017, p. 113 - 128
(RECOMMENDED CITATION)
DOI:10.24193/subbphilo.2017.2.08

THE CALCULATION OF CUT OFF SCORE IN ROMANIAN AS A FOREIGN LANGUAGE PRETEST AND EXAMINATIONS

DINA VÎLCU[1]

ABSTRACT. ***The Calculation of Cut off Score in Romanian as a Foreign Language Pretest and Examinations.*** While our educational and professional lives have been deeply influenced by our language abilities, companies and institutions rely more and more on formal evidence of these. Thus, organisations issuing language certificates have a great responsibility, related to the quality of the tests that they apply and the validity and reliability of the examinations. An essential element in this whole process is the cut off score, which needs to be set in relation with the purpose of the examination, the structure and the input of the test and the characteristics of the test taking population. The calculation of the cut off score in relation with these factors became central to the examination process for Romanian as a foreign language at Babeș-Bolyai University. The way in which this is set was decided in preparation for the auditing process with ALTE (the Association of Language Testers in Europe), an audit which our suit of examinations passed successfully. We apply the method of contrastive groups, based on the expertise of the test takers' teachers and on the students' results for every component of the examination and of pretest. Thus, the limit of passing/failing an examination is not set beforehand and just applied to each examination, all relevant factors being, instead, considered. The procedure we implemented for the calculation resulted in a higher reliability of the test takers' results and also of the tests in the long term.

Keywords: *language examinations, pretest, test component, cut off score, contrasting groups, test takers, bio data, data processing.*

REZUMAT. ***Calcularea scorului cut off în pretestările și examenele de română ca limbă străină.*** Parcursul educaţional şi viaţa profesională ne sunt profund influenţate de abilităţile noastre lingvistice, în acelaşi timp, diverse companii şi instituţii bazându-se tot mai mult pe dovezi formale ale acestora. Organizaţiile care emit certificate lingvistice au, astfel, o responsabilitate foarte mare, legată de calitatea testelor pe care le aplică şi de validitatea şi fiabilitatea examenelor. Un element esenţial în tot acest proces este scorul *cut off*, care trebuie să fie stabilit în

[1] Lecturer, Ph.D, at the Department of Romanian language, culture and civilisation, Faculty of Letters, Babeș-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

relație cu scopul examinării, cu structura și cu inputul testului, ca și cu caracteristicile candidaților. Calcularea scorului *cut off* în relație cu acești factori a devenit un element central al procesului de examinare a românei ca limbă străină la Universitatea Babeș-Bolyai. Felul în care scorul *cut off* este stabilit a fost decis în cadrul procesului de pregătire pentru auditarea examenelor noastre de către Asociația Testatorilor de Limbi din Europa (ALTE), un audit pe care seria noastră de examene l-a trecut cu succes. Aplicăm metoda grupurilor contrastive, care se bazează pe expertiza profesorilor ce le predau candidaților, ca și pe rezultatele studenților obținute pentru fiecare component al pretestării sau al examinării. Astfel, limita la care un anumit examen se trece sau se pică nu este stabilită dinainte și aplicată fiecărei examinări, toți factorii relevanți fiind, în schimb, luați în considerare. Procedura pe care am implementat-o pentru a calcula scorul *cut off* are ca rezultat o mai mare fiabilitate a rezultatelor obținute de către candidați și, de asemenea, a examenelor pe termen lung.

Cuvinte-cheie: *examene de competență lingvistică, pretestare, component al testului, scor cut off, grupuri contrastive, candidați, date personale, procesare de date.*

1. The relevance and value of cut off score calculation

The essential role of correctly determining the cut off score in a language examination can only be understood in direct relationship with the importance that language tests and their results have received in our educational, professional and social lives. In so many cases, test takers sit for an examination because the certificate they might obtain is necessary for their studies, for their job or for a dramatic change in their life, like moving to a foreign country. In these conditions, the responsibility of the test providers cannot be enough emphasized (Bachman, Palmer 1996: 99; Principles of good practice: 5; Spolski 2008: 299).

According to the Manual for *relating language examinations to the CEFR* (*Common European Framework of Reference for Languages: Learning, Teaching, Assessment*), “the crucial point in the process of linking an examination to the CEFR is the establishment of a decision rule to allocate students to one of the CEFR levels on the basis of their performance in the examination” (p. 11). A formal representation of this decision rule is given in the cut off scores, which define, according to Kane (Kane et al. 1999: 6) “the boundary between adjacent performance categories”. Anticipating the presentation of the way in which the cut off score is calculated for the examinations of Romanian as a foreign language (RFL) at the Department of Romanian language, culture and civilization, Babeș-Bolyai University, I will specify that in the case of our suite of examinations the cut off score we need to determine will make the difference between pass or fail within the level, the examinations being already conceived in accordance with and corresponding to the CEFR levels A1 – B2.

The limitations which a state educational system can impose on the calculation and decision over the cut off score might often function as a rigid point of reference, failing to consider a number of variables which result from the possible - or rather probable - differences from a class of test takers to the next in terms of population characteristics, syllabus followed, teaching methods, priorities, etc. The application of an inflexible, often arbitrarily pre-established cut off score (5 out of 10 in the Romanian system) might impact negatively on the reliability of the test, on decisions taken based on provided results and on the validity of the whole testing procedure. However, even functioning in the interior of an official educational system such as the Romanian one, the limitations mentioned above can be accommodated through the choice and the rigorous application of a method for setting the cut off score, followed by a method of corroborating the calculated cut off score with the official pass mark, a process which will be demonstrated further in this study.

As the literature in the domain reveals, the way to decide upon the cut off score for an examination is to get through a process of standard setting, by choosing one of the available methods and following its procedures. As Kaftandjieva points out (Kaftandjieva 2010: 10) "standard setting does not exist objectively. /…/ Standard setting is simply a summary of interpretations, opinions, beliefs of one or more individuals, authorized for one reason or another (professionalism, official position or at random), to serve to set the cut scores, which at a later point will be used as the basis for interpreting the test results." Since the process of standard setting has a profound subjective dimension (Horn et al. 2000: 3; Cizek 2001: 266; Kaftandjieva 2010: 10), what we need, in order to make it as reliable as possible, is the "proper following of a prescribed, rational system of rules or procedures" (Cizek 1993: 100) or, in other words, "well-conceived methods for setting performance standards, implementing those methods faithfully, and gathering sound evidence regarding the validity of the process and the result" (Cizek 2004: 46-47 apud Kaftanjieva 2010: 10). The choice of a particular standard setting method (in our case, contrasting groups), the selection of the participants in the panel of experts, the strict or justifiably adapted implementation of the indicated proceedings are all steps in the process and will be shortly presented in this study.

2. Method and methodology

2.1. Requirements of the standard setting method

Contrasting groups is an examinee-centred standard setting method. In order to apply it, two types of information about each test taker are necessary:

- the person's test score;
- a judgment of the adequacy of the test taker's knowledge and skills (Livingston & Zieky 1982: 31).

According to Livingston & Zieky (1982: 31), the judgments used in this method should meet the requirements below, the presentation of which is completed with the modalities in which we answered them in the case of the RFL examinations:

- *The judgments must be made by persons who are qualified to make them.* In our case, the persons selected for being part of the panel of experts are teachers and collaborators of the Department, specialized in teaching and assessing Romanian as a foreign language. Besides being very familiar with the examinees' linguistic knowledge and skills, the experts need to understand extremely well the level of the test which the candidates take. These features are vital for the adequate setting of the cut off score for a particular examination and the effectual accomplishment of this purpose depends on both aspects in equal measure.
- *The judgments must concern the knowledge and skills which the test is intended to measure.* Answering this requirement is quite challenging. In Livingston & Zieky (1982: 33) we find the example of teachers who, being asked to judge their students' skills in English composition, might let themselves influenced by the students' understanding of literature, for instance. In our case, the session of standard setting is organized for each of the skills and components which are part of the pretest and of the examinations (listening, reading, speaking, writing, elements of communication construction), for the levels A1-B2. This helps the experts focus on one particular skill at a time, without being distracted or influenced by various factors. We consider that the organization of the standard setting session for each component of the examination separately instead of the exam as a whole contributes to answering this second requirement of the process.
- *The judgments must reflect the test takers' skills at the time of testing*. Only in this way the estimation made by the members of the panel will justifiably corroborate with the live exam results obtained by the test takers. Our sessions of standard setting are always organized prior to the administration of each pretest and examination. This process is usually accomplished just days before the pretest and the examination, ensuring the appreciation of the test takers' skills at the time of testing.
- We consider as a way to validate our results the calculation of the cut off score also for the process of pretesting. The pretesting is applied to the same population (with a possible but not large variance in number) only days before the actual examination (needless to say, with the use of totally different tasks and items from the ones administered as part of the examination). The procedure of standard setting is applied twice, first in comparison with the results from pretesting, then in comparison with the results from the examination, for two reasons: 1) in order to have an

estimation of the cut off score before the tasks are introduced, in the next examination session (during one of the following academic years), in a live exam; 2) in order to be able to compare the results obtained after pretesting in one year with the results obtained after examination in one of the following years, so that the consistency of the cut off score can be confirmed.

- *The judgments must reflect the judges' true opinion*. Livingston & Zieky (1982: 33) give the example of teachers whose judgment might be influenced by different factors, like the idea that their estimation could be used to judge the value of their teaching. This is why the judges need to be explained as well as possible the purpose of standard setting and their essential contribution to the successful accomplishment of the process.

2.2. Standardisation

The above presentation of the requirements concerning the judgments involved in the process of standard setting through the method of contrastive groups reveals also how important the role of the experts who are members in the panel is in the whole process. The selected experts also need to understand the responsibility of their involvement in the standard setting process and, before taking a final decision on the use of this method, both the coordinator of the process and the experts selected need to be sure that they meet the corresponding requirements. As far as the RFL tests are concerned, the experts selected for the standard setting process are characterized by the following features:

- they are qualified and/or experienced persons in teaching and assessing Romanian as a foreign language;
- they are familiar both with the content standards and with the performance standards involved by these examinations and the standard setting session;
- they are familiar, in a certain measure, with the CEFR and the whole RFL teaching and testing system we built based on it;
- they know very well the level of competence, the abilities and the skills specific to the students at the moment they take the test;
- they willingly participate in standard setting sessions and consider that this is a good exercise for themselves and a welcome addition to the testing procedures we apply;
- every time we apply this procedure we make sure that they are in sufficient numbers to provide reasonable assurance that the results would not vary greatly if the process were replicated;
- they are open to discussion and ready to provide valuable feedback on the procedure of standard setting itself and on the materials used during workshops.

A complete process of standard setting could include the stages presented below. It is recommended to develop this set of operations periodically, according to the needs of the experts involved. However, even if not all the stages are developed every time the standard setting process is applied, the coordinator of this activity must adapt and offer the experts the phases they need to get through.

The members of the panel are usually familiar with the CEFR in different degrees, so organising a complete session of familiarisation, as recommended by the Manual, might be indicated, in order to bring everyone, as far as it is possible, to the same or to a very similar understanding of the CEFR and of the assessment instruments. The stages presented below and developed as part of the standard setting process in our department follow quite closely the structure indicated in the *Manual of relating examinations to the CEFR* (p. 35-57).

2.2.1. Preparation for the workshop. The preparation for the familiarization workshop can begin, for reasons of time save and also of pre-familiarisation, before the actual workshop is developed, for example, with some homework. In our case, the participants in the panel were kindly asked to do, before the workshop, the following tasks:

- The participants had to read Section 3.6 from CEFR (*Content coherence in Common Reference Levels*) and Table 1 from the CEFR (*Common Reference Levels: global scale*) and to identify/mark in Table 1 the features which characterise each level, differentiating it from the adjacent ones.
- Having in view the main category of students they work with (the students in the preparatory year), the level B2 and the most frequent situations in which they thought the students would have to use Romanian, the participants needed to answer questions focused on the main components addressed in the workshop: reading and listening (What kind of texts will the students have to read/listen to? What is the purpose for which they will have to read/listen to these texts? Which are the main abilities the students/candidates need to possess in order to successfully accomplish the reading/listening activities?).

The workshop can open with a group discussion on the homework, during which truly interesting ideas and conclusions will most certainly emerge.

2.2.2. Familiarisation with the CEFR. The stage of familiarization included the following activities:

- Self-assessment in two foreign languages. Each expert assessed their own linguistic abilities in two foreign languages. They used Table 2 in the CEFR (*Common Reference Levels: self-assessment grid*) and the abilities to consider were: listening, reading, oral interaction, oral production and writing. The activity was done individually, the results being subsequently discussed with the whole group. Many interesting characteristics were discovered and discussed. Some ideas which emerged were: the experts often did not choose the foreign language they spoke the best for self-assessment; there were big differences between the abilities manifested for the same language (always a matter of discussion while assessing candidates); the context in which the language had been learned proved to be extremely relevant; the specialization of the participants in teaching and assessing languages became evident due to some specific points they made (e.g. I am at B2 level in writing in French as long as I am offered the format of the document I need to produce – formal letter, e-mail, etc.).

- Receptive skills. Identification of salient characteristics for the levels A1-B2. Table A2 from the *Manual* (*Salient characteristics: Reception*) was used in order to identify the main features the receptive skills are defined through at the levels A1-B2 (setting, action, what is understood, source, restrictions). While the table also contains level C1, we usually focus on the levels A1-B2, since they are the ones we usually assess with our candidates.
- Qualitative analysis of the CEFR Scales. Listening, Reading, Elements of communication construction. For Listening, two of the illustrative scales provided in the CEFR were used for the workshop: *Overall Listening Comprehension* and *Listening as a member of live audience*. The descriptors were mixed and the participants needed to arrange them in the correct order. The activity was then repeated for Reading, with the CEFR illustrative scales *Overall reading comprehension* and *Reading for information and argument*. We usually choose these scales because they match the most frequent situations in which our candidates would need to listen to spoken productions or to read texts. For Elements of communication construction the illustrative scales *Vocabulary range*, *Vocabulary control* and *Grammatical accuracy* from the CEFR were chosen. After each activity of qualitative analysis of the scales, the results were discussed with the whole group and the essential, most relevant characteristics in each descriptor were identified.
- Reconstruction of the local grids. The set of instruments we use at the Department as a theoretical support for creating tasks and items includes grids for the receptive skills and for the linguistic competence, which we developed based on the illustrative scales in the CEFR. The reading grid has the following components: general description, tasks and strategies, types of texts and contexts, characteristics of the written text. The descriptors were elaborated for the levels A1-B2. The benefit of using this kind of local instruments during the workshop comes from the fact that

the participants become more familiar with the structure and the characteristics of the tests applied locally, having, at the same time, the possibility of correlating them with the CEFR. A group discussion following the use of these instruments can reveal appreciation of the participants and suggestions for the improvement of the material. During the workshop presented here, the following points were discussed:

- general description, level B1: how exactly we could incorporate in items elements which could reveal the fact that the candidate can identify the arguments and can understand the main conclusion in clearly structured argumentative texts;
- tasks and strategies – descriptors to add to the grid: level A1: the test takers' capacity of identifying anaphoric and cataphoric references; the capacity of understanding detailed information; level B1: the capacity of summarising information for revealing a conclusion;
- types of texts and contexts, level B2: to include the text of a scientific nature;
- characteristics of the written text, level A1: to include the use of punctuation.

The activity was repeated for Listening. The categories in our grid are: general description, tasks and strategies, types of texts and contexts, characteristics of the recorded text.

In order to re-familiarise with the content regularly tested within the component Elements of communication construction, the participants were presented with the detailed specifications of grammar and vocabulary of the examinations.

2.2.3. Training for standard setting. Illustration. In preparation of the illustrative stage of the training, the *Content Analysis Grids* in the *Manual* (Section B1: *CEFR Content Analysis Grid for Listening & Reading*) was used in order to present and describe samples of the Reading and Listening components. In our case, information on the tasks, input texts and items in the examinations are usually provided. A document is assembled for illustrating the level or levels which need to be illustrated. The participants are given time to study the task samples and the grid and to discuss the elements which characterise the component at each level. As for all the other stages of the process of preparation for setting the cut off score, we normally try to relate every activity to the examinations our test takers sit for and to their specific needs in relation to subsequent use of Romanian as a foreign language. Once again, the worksheets used for this stage give the experts the occasion to discuss the contexts the test takers need to show ability in, the communication themes they are expected to meet and get involved with, the communicative tasks, activities and strategies and also the text types the candidates are expected to be able to handle.

Unlike other languages (English, French, German, etc.), no exemplary samples for different skills and levels are available for Romanian. Following consultation with experts, we provided samples in the form of a text with the corresponding task and items.

As indicated in the *Manual* (p. 49), the activity started with the actual solving of the tasks by the participants, whose answers were confronted with the key. The participants then gave their estimation of item difficulty. Their estimations were registered and compared and the differences which resulted were discussed and clarified.

2.2.4. Training for standard setting. Controlled practice and individual assessment. These two stages will be presented here together because the activities they included are the same, the difference being given by the involvement of the coordinator. The worksheets prepared for these stages of the training for standard setting contained: the task (in this case, multiple choice), the input text, the items, a box with numbers of the items for the experts to write their answers, and an adaptation of Form C5 from the *Manual*: a table with four columns (1. the number of each item, 2. the CEFR level the expert attributed to the item, 3. the operationalised descriptor and 4. observations – this last column was meant for notices related to the quality of the item, its clarity, its connection with the part in the text it relates to in order to be solved, etc.).

The sequence of activities and the principles applied for working on every worksheet were as follows:

- The experts were allowed to use plus levels. Even if we do not include in our tests tasks or items for plus levels, the experts could use them in the appreciation of the items first of all because they sometimes felt they could not decide for one or the other of the levels, but also because we considered that the results of the analyses would be more relevant and closer to the real appreciation of the experts.
- For each component approached here (reading, elements of communication construction and listening), there were three phases, organised according to the instructions in the *Manual*: Illustration (for which we chose an exemplary task from the item bank); Controlled practice (for which we chose 4 tasks, one for each level we were interested in, without any mentioning of this approach to the experts) and Individual Assessment (for which we chose three tasks of different levels).
- The task was first solved by the experts individually and then their answers were compared to the key provided by the coordinator.
- The experts attributed individually a level to each item of the task. The options of the experts were written in a table and compared. For each item, a difference in appreciation of two levels was permitted for the first

two phases (Illustration and Controlled Practice). Anyway, the differences in the level attributed were discussed and explained. For Phase 3 (Individual Assessment), according to the indications in the *Manual* (p. 50), a spread of results of no more than one and a half levels was targeted. In case there were greater differences, the results were discussed with the whole group and a new round organised for the task. Then the results were represented in graphs. Each interval in the graph represents half a level: 0,5 – below A1; 1 – A1; 1,5 – A1/A2; 2 – A2; 2,5 – A2/B1; 3 – B1; 3,5 – B1/B2; 4 – B2; 4,5 – B2/C1.

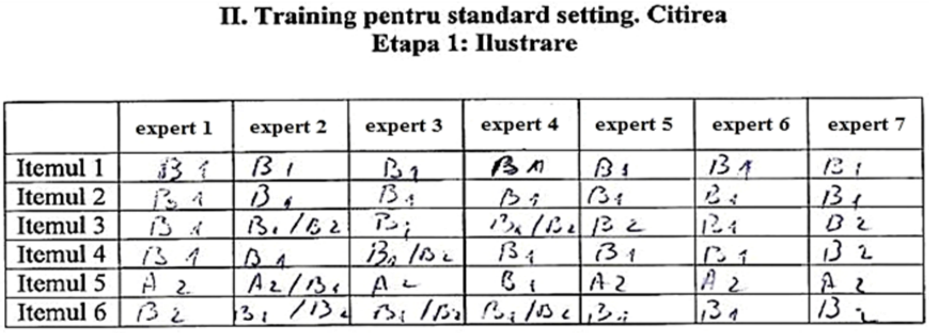
- The experts tried to figure out the operationalised descriptor for each of the items, following the categories in the descriptive grid (general description, tasks and strategies, types of texts and contexts, characteristics of the written text).
- The experts wrote observations on the task and the items.
- There was a group discussion on the results.

2.2.5. Example. In case of Worksheet 13 (Illustration: Reading – Multiple choice; level B1), the results were as follows (Table 1 and Graph 1 below):

- Solving the task: There was one mistake made by one expert, due to lack of attention, in item 6; no discussions were necessary on the matter.
- Level attributed to the items: Maximum difference among the experts: one level for four of the items;
- The operationalised descriptors were identified in relation with the grid for Reading developed at the Department;
- Some of the qualitative observations related to the items were: item 2: the tense of the sequence of sentences should be present, not past, to be closer to the input text; item 3: would be better replaced because the correct answer depends on the answer to item 2; there were also general observations related to descriptors possible to be added in the descriptive grid (expressing option, consequence; rephrasing; paraphrasing; identifying reference; processing of scientific text).

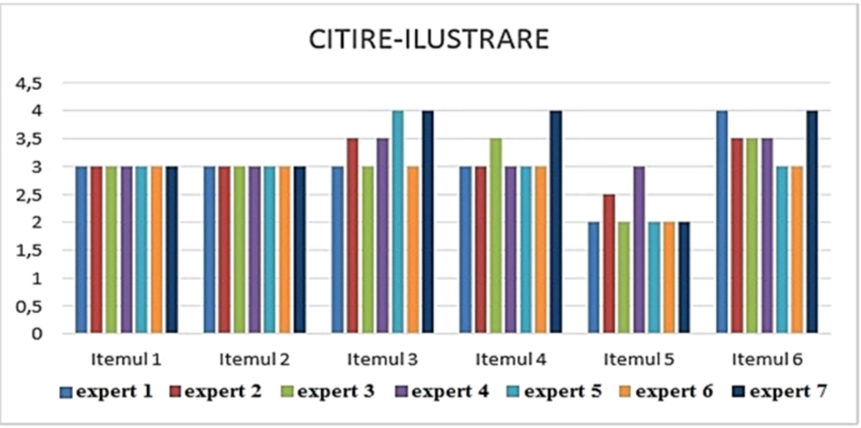
Once the standardization process is closed, it can be considered that all the members of the panel have the same or a very similar understanding of the levels in the CEFR, of the structure and objectives of the test tasks and also of the process of item writing. The following stage, of standard setting, can now begin. As the real target of the whole process described so far in this study, it needs to be organized very carefully, starting with the preparation of the materials and finishing with the registration and processing of the results. The standard setting method which we chose to use (the contrasting groups method) will be presented below, together with considerations related to the particular case of our examinations.

Table 1 and Graph 1



II. Training pentru standard setting. Citirea
Etapa 1: Ilustrare

	expert 1	expert 2	expert 3	expert 4	expert 5	expert 6	expert 7
Itemul 1	B 1	B 1	B 1	B 1	B 1	B 1	B 1
Itemul 2	B 1	B 1	B 1	B 1	B 1	B 1	B 1
Itemul 3	B 1	B1 / B 2	B 1	B1 / B2	B 2	B 1	B 2
Itemul 4	B 1	B 1	B1 / B2	B 1	B 1	B 1	B 2
Itemul 5	A 2	A 2 / B1	A 2	B 1	A 2	A 2	A 2
Itemul 6	B 2	B1 / B2	B1 / B2	B1 / B2	B 1	B 1	B 2



2.3. Standard setting

The contrasting groups method is based on the idea of dividing the test takers into two groups, one of which will be considered qualified and the other unqualified. The decision will be taken on the basis of judgments of their knowledge and skills, made by a panel of experts. Livingston & Zieky (p. 35) consider that a choice for determining the passing score would be to place the limit between the two categories at the point where there are just as many qualified test takers as unqualified. This solution would be adequate especially if examinees from all the score range possible for the test are selected. After the procedure has been performed, the proponents of this method recommend a stage of "smoothing" of the data (p. 36-40), since the percentage will not increase steadily from one level to the next. However, we chose to adopt a more reliable procedure from the start, the one recommended in the *Manual* (p. 67-68), of constructing decision tables for several cut-off scores.

Table 6.4 below, from the *Manual* (p. 68) shows an example for setting the standard for the levels B1 (or lower) and B2 (or higher), for a group of 400 students, in relation with a test containing 50 items. The low scores (up to 20) and high scores (from 28 on) are taken together; the other scores are displayed separately, these representing the possible cut off scores which will be calculated (Table 6.5 below).

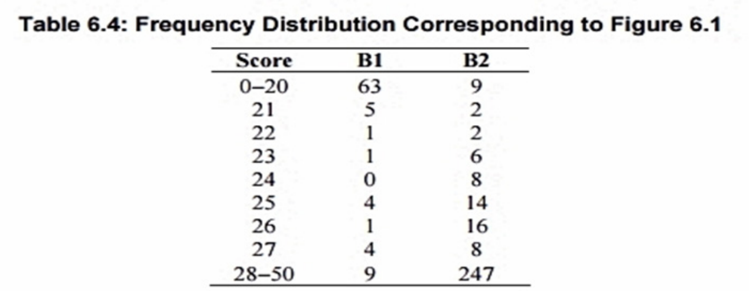


Table 6.4: Frequency Distribution Corresponding to Figure 6.1

Score	B1	B2
0–20	63	9
21	5	2
22	1	2
23	1	6
24	0	8
25	4	14
26	1	16
27	4	8
28–50	9	247

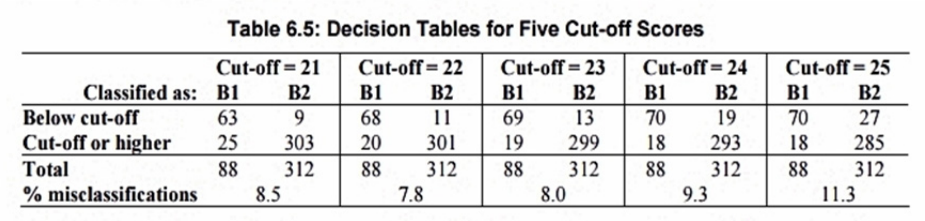


Table 6.5: Decision Tables for Five Cut-off Scores

	Cut-off = 21		Cut-off = 22		Cut-off = 23		Cut-off = 24		Cut-off = 25	
Classified as:	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
Below cut-off	63	9	68	11	69	13	70	19	70	27
Cut-off or higher	25	303	20	301	19	299	18	293	18	285
Total	88	312	88	312	88	312	88	312	88	312
% misclassifications	8.5		7.8		8.0		9.3		11.3	

For each cut off score the number of misclassified students is calculated and the one where the misclassified are the least numerous will be the cut off score. Thus, in the case presented above, according to Table 6.5, the cut off score will be 22.

Given the fact that we chose to calculate the cut off score for each test component and that we grant a total of 20 points for each of these components, we construct decision tables between the limits of 8 and 12. If the number of misclassified students is the same for more than one cut off score, resulting, thus, in two possible scores, we decided to choose the one which is the closest to 10, half the number of points which can be awarded for a test component. After calculating the cut off score for each component, in the case of the examinations we aggregate the results and obtain the cut off score of the examination for that particular test and for that session of administration, adapting, thus, as much as possible to the test taking population through the consideration of all the elements involved: basically the structure and estimated difficulty of the items, on the one hand, and the characteristics of the test takers, on the other hand. The cut off score is situated at one of the values between 46 and 54, which is equivalent to 5, the mark specified in the law of education as the passing mark for an examination in the Romanian educational system. We do not aggregate the cut off scores from all the components in the case of pretesting because the components applied in pretesting might not be administered in exactly the same combination for a future live examination. Their items are submitted to modifications after the results obtained by the candidates are statistically processed, and included into an item bank from which they will be selected for live examinations.

An element which increases reliability of the method we chose is the fact that we include in the calculation of the cut off score not only a representative proportion of the test takers, but actually all of the participants in the case of pretesting and at least 90% in the case of examinations. This representativeness is made possible by the fact that currently we have a rather reduced number of test takers for every session of pretesting and of examination (about 120). This aspect is seen, from a statistical point of view, as a shortcoming, since the relevance of statistical calculations grows with the number of subjects involved. In the case of our examinations, the rather reduced number of participants allows us to involve all or almost all of them in our calculations, maximizing, thus, the relevance of the results and, consequently, of the cut off score.

2.4. Advantages and disadvantages of the contrastive groups method

"There is no one method that is best for all testing situations" (Livingston & Zieky 1982: 53). Judging the particular case of our suite of examinations, the application of the contrasting groups method has advantages and disadvantages. They come from the characteristics of the method itself, but also from the specific situation of our examinations. The advantages of using this method in general and also in the case of our examinations in particular are the following:

- "...people in our society are accustomed to judging other people's skills as adequate or inadequate for some purposes – especially in educational and occupations settings /.../ therefore, making this type of judgment is likely to be a familiar and meaningful task." (Livingston & Zieky 1982: 31); the teachers and collaborators of our department, involved as experts in the standard setting procedure, are particularly accustomed to making judgments on the knowledge and skills of their students, since they need to assess them periodically during the academic year, to give them feedback and to adjust the educational process according to results;
- we had the chance of using the test scores of real test-takers (which cannot be applied for all methods of cut off score calculation): for receptive skills and for elements of communication construction (grammar and vocabulary) we used also the results from pretesting; we could be reasonably sure that the judges would base their judgments on the same qualities of the test takers that the test measures; for productive skills we used results from real examinations, being able to present to the

experts samples of written productions for a large part of the participants in the examinations; the experts being teachers of the candidates whose results were used for the process of standard setting, they had direct access to the students' written productions and were familiar with their capacity of producing oral discourse (Cf. Livingston & Zieky 1982: 53);

- the cut off score derived using this method has "a high degree of stability and adequacy" (Kaftandjieva 2010: 34);
- in the case of our examinations, the experts (one or more) knew very well the capacity of the specific examinees who were classified.

The disadvantages of applying this method and also the modalities in which we tried to compensate for them come as follows:

- the specialized literature registers common practical difficulties in applying this method, like bringing together teachers of students from different centres, schools, different institutions; as far as our examinations were concerned, this was not the case, given the fact that the students whose results were subjected to the standard setting procedure study in the same institution, the Faculty of Letters, and the fact that the teachers are members or collaborators of the same Department;
- the number of judges is usually limited to one for each examinee (Cf. Kaftandjieva 2010: 34); however, in our case, this did not apply – for each examinee, at least two experts and more often more than two experts could emit judgments, have a discussion and agree on a verdict; given the structure of the academic programme the students are enrolled in (our students have 25 hours of general Romanian course per week in the first semester and 15 hours in the second semester, completed with classes of specialized language and culture and civilisation), each group has normally two to four teachers, who get to know the students in a consistent measure; in plus, there is another teacher (or sometimes two teachers per semester) who replaces teachers when they are not available for classes; this person (these persons) also offer(s) extra tuition for students whose results in studying are not as expected to be at certain moments during the course; moreover, they can sometimes be part of the assessment commissions for oral examinations; thus, one, two or three extra opinions are available for deciding on each student during the standard setting session, the number of experts expressing judgments for every student being, normally, between 2 and 4; in case the experts cannot agree on a verdict concerning a student, the main teacher(s) of the group the student is part of emits the final opinion.

Every stage of a language examination process needs to be validated, starting with the test construction and finishing with the standard setting. According to the *Manual* (p. 90) "Validation concerns the body of evidence put forward to convince the test users that the whole process and its outcomes are trustworthy." We considered validation very carefully in relation with every stage of the testing process and, as standard setting is a crucial part of the whole testing process, we chose to provide evidence for every step on the way. Without entering into details here, I will just specify that the stage of standardization for cut off score was validated from different points of view: explicitness, practicability and implementation. Intra-judge consistency and inter-judge agreement were also calculated.

3. Conclusions

The procedure of standard setting proved very useful for the process of developing and administering examinations at the Department of Romanian language, culture and civilization. The advantages of applying this procedure were multiple and some of them are presented below:

- The most important advantage of the procedure of standard setting is the contribution it has for the quality, validity and reliability of our examinations. Besides offering the basis for cut off score in the case of each examination we apply, this procedure represents a supplementary filter which can confirm the value of our tasks and items or can reveal problems in their quality.
- The session of standard setting proved to be an extremely welcome occasion for the members of the expert panel, employees or collaborators of the Department of Romanian language, culture and civilization, to refresh and improve their knowledge of the Common European Framework of Reference and related documents. This was a very helpful exercise of reflecting on our examinations and proposing solutions for making them better. All the participants were involved in the process, there were numerous discussions on various topics (which determined us to extend the duration of the session with one more meeting) and many valuable suggestions were made.
- The procedure of standard setting was an occasion to discuss with our colleagues newly designed instruments describing the components of listening and reading (illustrative grids with descriptors) and to get feedback on them.
- The session of standard setting represented an occasion for every teacher to reflect on the relationship between the examinations and their teaching methods. This even had implications on the organization of the courses in our Department. The students who came late in the preparatory year and missed a portion of the teaching programme were the most problematic to evaluate, their performance being most of the times extremely fluctuant and unpredictable. We concluded that

situations like these, in which the students need to take an examination without having a sufficient number of hours of study should definitely be avoided and we took measures to accomplish this.

- The first time when we applied the procedure of standard setting, on the occasion of the preparation for the ALTE audit, gave us the opportunity to learn a lot of new things related to assessment, including at the level of data processing, which proved extremely useful in our activity.

The procedure of standard setting became current practice for each of our examinations, proving to be an instrument of great value in the process of quality assurance for our students and bringing more certainty in relation to our work of creating and developing examinations and assessing candidates' language abilities for ourselves.

BIBLIOGRAPHY

Bachman, Lyle F., Palmer, Adrian S., *Language Testing in Practice. Designing and Developing Useful Language Tests*, Oxford, Oxford University Press, 1996.

Cizek, Gregory (Ed.), 2001, *Setting performance standards: Theory and applications*, Routledge.

Cizek, Gregory, 1993, "Reconsidering Standards and Criteria", in *Journal of Educational Measurement*, Vol. 30, Issue 2, p. 93-106.

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

Horn et al., *Cut scores: results may vary*, 2000, NBETPP, Monographs, vol I, No. 1, 2000, p. 1-31.

Kane, Michael, Crooks, Terence, Cohen, Allan, "Validating Measures of Performance", in *Educational Measurement. Issues and Practice*, Vol. 18, Issue 2, 1999, p. 5-17.

Kaftandjieva, Felianka, *Methods for Setting Cut Scores in Criterion-referenced Achievement Tests. A Comparative analysis of six recent methods with an application to tests of reading in EFL,* CITO, Arnhem, 2010.

Livingston, Samuel A, Zieky, Michael J, 1982, *Passing scores: A Manual for Setting Standards on Performance on Educational and Occupational Tests*, Educational Testing Service.

Principles of Good Practice for ALTE Examinations. Revised draft October 2001.
(http://www.alte.org/attachments/files/good_practice.pdf)

Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual, 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

Spolski, Bernard, "Introduction. Language Testing at 25: Maturity and responsibility?" in *Language Testing* 2008, 25 (3), p. 297-305.